UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

REGRESIONES APARENTEMENTE NO RELACIONADAS:
UNA PERSPECTIVA BAYESIANA

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

ACTUARIA

P R E S E N T A :

ZINNYA DEL VILLAR ISLAS

Director de Tesis: Dr. Eduardo Gutiérrez Peña

2008

A mi madre

AGRADECIMIENTOS

A la UNAM, en especial a la Facultad de Ciencias y al IIMAS por albergarme en sus aulas y por ofrecerme una educación de alta calidad.

Al Dr. Eduardo Gutiérrez Peña por brindarme el espacio, respaldo y tutoría de este trabajo durante los años en que lo llevé a cabo.

A las profesoras de la Facultad de Ciencias, Margarita Chávez Cano y Ruth Fuentes García por sus valiosos comentarios en la elaboración de este trabajo.

A Nico por la confianza, comprensión y apoyo en la concusión de este trabajo, que es solo un paso para el futuro que esperamos.

A Fabi y Chucho por estar siempre a mi lado e impulsarme a mi superación día con día. Los buenos tiempos ya están con nosotros de nuevo y esto se debe también gracias a Ger, Emilio, Bruno y Carlos.

A mis grandes amigos que estuvieron en las buenas y en las malas que se vivieron durante la elaboración de este trabajo, los quiero.

Regresiones aparentemente no relacionadas: una perspectiva bayesiana

Zinnya del Villar

Agosto, 2008

Índice

1. Introducción

1.1. Introducción

Las técnicas de regresión lineal son de los métodos estadísticos más utilizados para conocer la relación existente entre dos o más conjuntos de datos observados. La pregunta es cómo varía una variable y como función de otra variable o conjunto de variables z. Es decir, nos interesa la distribución condicional de y dado z, parametrizado como $p(y|\theta, z)$, donde y es llamada variable de respuesta, mientras que $z = (z_1, ..., z_k)'$ se conoce como variable explicativa.

Dentro de los modelos de regresión lineal, existe aquel que puede ser escrito como p regresiones múltiples lineales separadas, cuyas variables de respuesta están relacionadas entre sí, aún cuando superficialmente no lo parezca; es decir, cada ecuación estima la variable de respuesta respectiva con diferentes variables explicativas. Si las ecuaciones están usando los mismos datos, los errores tal vez estén correlacionados a través de las ecuaciones. Por supuesto, resulta más eficiente estimar las ecuaciones conjuntamente que cada una por separado. Dicho modelo es conocido como SUR, Regresiones Aparentemente no Relacionadas, que en inglés sus siglas se leen como *Seemingly Unrelated Regressions* y lo podemos ver como una extensión de los modelos de regresión lineal que permite errores correlacionados entre las ecuaciones.

El modelo SUR fue analizado y publicado por primera vez por Arnold Zellner en Zellner (1962) como una técnica econométrica para analizar un sistema de ecuaciones múltiples con parámetros de restricción y errores correlacionados. En algunos casos el uso de regresiones aparentemente no relacionadas puede verse como un método de unión de series de tiempo; sin embargo, la distinción recae en la correlación de los errores y en el supuesto de que cada unidad tiene un vector de coeficientes diferentes. En el capítulo de preliminares se discutirá detalladamente el modelo SUR.

El propósito de este trabajo consiste en describir y analizar dicho modelo desde una perspectiva bayesiana, ya que la densidad final de los parámetros del modelo de regresiones aparentemente no relacionadas no puede, en general, ser evaluada analíticamente, por lo tanto se propone el muestreo Gibbs como aproximación. Esta aproximación será analizada posteriormente con datos reales en un aplicación econométrica del modelo SUR.

1.2. Mapa de tesis

El capítulo 2 describe las técnicas estadísticas y computacionales utilizadas en el desarrollo del presente trabajo, tales como, inferencia bayesiana, regresión lineal multivariada, la descripción del modelo de regresiones aparentemente no relacionadas y los métodos computacionales, que incluyen la integración Monte

Carlo vía cadenas de Markov y el muestreo Gibbs; así como el software utilizado, S-Plus, R, y WinBUGS.

El capítulo 3 describe con detalle el modelo de regresiones aparentemente no relacionadas desde la perspectiva bayesiana y el uso de los métodos computacionales del muestreo Gibbs para aproximar la densidad final.

Algunas aplicaciones se mencionan en el capítulo 4, y con detalle se analiza la aplicación del modelo SUR a un análisis econométrico con datos reales de México. Los resultados obtenidos se comparan con las diferentes perspectivas.

La conclusión sobre el uso de la aproximación Gibbs a partir de la perspectiva bayesiana contra la perspectiva clásica está dada en el capítulo 5.

1.3. Notación

Mientras sea posible, la misma notación se mantendrá en el desarrollo del trabajo. Los datos observados serán denotados con letras romanas $x, y, ...$, mientras que los datos no observados con $\tilde{x}, \tilde{y}, ...$ o con mayúsculas para algunos casos; los parámetros con letras griegas $\theta, \phi, ...$ Los escalares se denotan con minúsculas, los vectores con minúsculas en negritas y las matrices con mayúsculas en negritas. Los vectores siempre se acomodan por columna y la transpuesta de un vector $\boldsymbol{x}$ es denotada por $\boldsymbol{x}'$. La probabilidad de un evento A, $p(A)$, el valor esperado y varianza $E(A)$ y $Var(A)$, covarianza y correlación, $Cov(A)$ y $Corr(A)$.

2. Preliminares

2.1. Herramientas estadísticas

2.1.1. Inferencia bayesiana

La inferencia estadística, tanto clásica como bayesiana elabora conclusiones o inferencias sobre un parámetro θ o sobre los datos no observados $\tilde{y}$, condicionados en los valores observados de y; es decir, $p(\theta|y)$ o $p(\tilde{y}|y)$. Es en este punto de condicionamiento de datos, donde el enfoque bayesiano y el frecuentista divergen, ya que el enfoque bayesiano incorpora formalmente el conocimiento previo de θ, a través de una densidad $p(\theta)$. A pesar de esta diferencia, en varios análisis simples pueden resultar conclusiones similares utilizando los dos enfoques.

El paso inicial en la inferencia bayesiana, es construir la función de densidad de probabilidad conjunta para θ y y, de los datos no observados dados los datos observados. Este conocimiento incorpora información previa del fenómeno en estudio y solo está basada en los valores de los datos observados (cuando éstos están disponibles). La densidad de probabilidad conjunta puede ser escrita como un producto de dos densidades, que se refieren a la distribución inicial $p(\theta)$ y a la distribución muestral $p(y|\theta)$, respectivamente:

$$p(\theta, y) = p(\theta)p(y|\theta). \tag{2.1}$$

Condicionando sobre los valores observados y y usando el Teorema de Bayes, se obtiene la densidad final:

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(\theta)p(y|\theta)}{p(y)} \tag{2.2}$$

donde,

$$p(y) = \sum_{\theta} p(\theta)p(y|\theta)$$

ó,

$$p(y) = \int p(\theta)p(y|\theta)d\theta \tag{2.3}$$

para el caso discreto o continuo, respectivamente.

Una forma equivalente de (2.2) omite el factor $p(y)$, el cual no depende de θ, y al fijar y puede ser considerada una constante; así se obtiene la densidad final no normalizada,

$$p(\theta|y) \propto p(\theta)p(y|\theta). \tag{2.4}$$

Esta simple expresión es el núcleo de la técnica de inferencia bayesiana, en donde lo principal es desarrollar el modelo $p(\theta, y)$ de la manera más apropiada.

Al trabajar con el Teorema de Bayes, las dificultades computacionales surgen desde el momento en que es necesario calcular la constante de proporcionalidad que aparece en el denominador:

$$p(y) = \int p(\theta)p(y|\theta)d\theta$$

Por otro lado, asignar una distribución inicial, en general, no es un problema sencillo; sin embargo, empezaremos considerando una familia de posibles distribuciones iniciales que de lugar a cálculos simples. Si dicha familia es suficientemente flexible, es decir, que cubra un rango amplio de formas posibles, podemos utilizar el elemento de la familia que describa de la mejor manera nuestro estado de conocimiento acerca del valor de θ. Así, eligiendo adecuadamente la forma de la distribución inicial, se obtendrá una distribución final que pertenezca a la misma familia que la distribución inicial. Una familia de distribuciones es conjugada si tanto la distribución inicial como la final pertenecen a ella. Esencialmente el único caso en el que se pueden construir fácilmente familias conjugadas es para modelos que pertenecen a la familia exponencial.

Si existe un estado de ignorancia inicial, o una situación en donde no se quiere o no se puede hacer uso de la información inicial disponible, se utilizan las distribuciones iniciales no informativas. Entre los bayesianos existe controversia sobre la especificación de éstas, ya que a menudo dicha especificación conduce a distribuciones impropias (distribuciones que no integran a 1 como está descrito por la teoría de probabilidad).

Una de las distribuciones iniciales no informativas más comunes es la definida por la *regla de Jeffreys*, dada bajo condiciones de regularidad por $p(\theta) \propto |I(\theta)|^{\frac{1}{2}}$ donde,

$$I(\theta) = E\left[-\frac{\partial^2 \log p(y|\theta)}{\partial\theta\partial\theta'}|\theta\right]$$

es la matriz de información esperada de Fisher para θ. Jeffreys propuso la forma de obtener distribuciones iniciales no informativas consistentes entre reparametrizaciones del modelo, es decir, invariantes ante transformaciones uno a uno (invariante en el sentido de que respeta la regla de cambio de variable). En general, esto conduce a densidades en la forma $p(\theta) \propto k$ para parámetros de localización θ, y $p(\sigma) \propto \sigma^{-1}$ para parámetros de escala σ. Cuando un parámetro de localización θ y uno de escala σ están presentes, Jeffreys sugiere un cambio sobre la regla del producto que conduce a $p(\theta, \sigma) \propto \sigma^{-1}$.

Otra definición de una distribución inicial no informativa es la dada por Bernardo (1979), basada en medidas de discrepancia esperada de información y bajo el supuesto de normalidad asintótica, y coincide con la inicial Jeffreys en el caso univariado. En el caso multidimensional, el método de referencia trabaja

a partir del vector de parámetros en grupos y parece evitar algunas dificultades de otras aproximaciones en el caso de multiparámetros.

Otro elemento importante de la inferencia bayesiana es la distribución marginal o predictiva de y con densidad $p(y)$ dada por (2.3), también conocida como distribución predictiva inicial; inicial porque no es condicionada en una observación previa del proceso y predictiva porque es la distribución para una cantidad que es observable. Ésta obtiene la distribución esperada para la observación y como $p(y) = E[p(y|\theta)]$ y la esperanza es tomada con respecto a la distribución inicial de θ. Una aplicación similar puede ser tomada para la predicción de futuras observaciones $\tilde{y}$ después de observar y. Esta predicción se basa en la distribución de $\tilde{y}|y$, esto es, en la descripción probabilística basada en la información disponible. Si $\tilde{y}$ y y son condicionalmente independientes dado θ, entonces:

$$p(\tilde{y}|y) = \int p(\tilde{y}, \theta|y)d\theta = \int p(\tilde{y}|\theta)p(\theta|y)d\theta$$

Si $y = (y_1, ..., y_n)'$ y $\tilde{y} = (\tilde{y}_{n+1}, ..., \tilde{y}_{n+m})'$ son muestras de $p(y|\theta)$, la distribución predictiva es entonces usada para predecir valores futuros de la variable en la población. En el caso de parámetros multivariados $\theta = (\theta_1, ..., \theta_d)'$, las distribuciones marginal y condicional de los componentes θ_i pueden ser obtenidas de la función de densidad conjunta $p(\theta_1, ..., \theta_d|y)$. Para cada θ_i la posible distribución condicional es

$$p(\theta_i|\theta_j, j \in C) = p(\theta_i, \theta_j, j \in C)/p(\theta_j, j \in C)$$

para toda $C \subset \{1, ..., i-1, i+1, ..., p\}$.

Al obtener la distribución final, se resume la información a través de medidas de localización y de dispersión: media, moda, mediana, varianza, desviación estándar, precisión, rango intercuartil y curvatura en la moda. Con excepción de la mediana, todas estas medidas pueden ser evaluadas para las distribuciones conjuntas, marginal y condicional. La mediana solo tiene sentido para distribuciones univariadas.

Cabe recalcar que cuando utilizamos el enfoque bayesiano, los datos y afectan a la inferencia final solo a través de la función $p(y|\theta)$, la cual cuando la consideramos como una función de θ para y fija, es llamada función de verosimilitud. De esta manera, la inferencia bayesiana obedece lo que algunas veces es llamado el principio de verosimilitud, el cual dice que cualesquiera dos modelos de probabilidad que tienen la misma función de verosimilitud producen la misma inferencia para θ.

Desafortunadamente la implementación de las técnicas bayesianas usualmente requiere de un esfuerzo computacional muy alto. La mayor parte de este esfuerzo se centra en el cálculo de ciertas características de la distribución final del parámetro de interés. Así, por ejemplo, para pasar de una distribución con-

junta a una colección de distribuciones y momentos marginales que sean útiles para hacer inferencias sobre subconjuntos de parámetros se requiere integrar.

2.1.2. Regresión lineal multivariada

Las técnicas de regresión son de los métodos más utilizados en la estadística aplicada. Dada una variable de respuesta Y y un conjunto de covariables $\boldsymbol{z} = (z_1, ..., z_r)'$, es de interés estimar una supuesta relación funcional entre Y y $\boldsymbol{z}$, así como predecir el valor de observaciones futuras para distintos valores de las covariables.

Una manera de modelar dicha relación consiste en representar el valor esperado de Y como

$$E(Y|\boldsymbol{z}) = \mu(\boldsymbol{z}),$$

donde, en general, $\mu(\cdot)$ es una función desconocida. En la práctica es común aproximar a $\mu(\cdot)$ a través de una función paramétrica simple,

$$\mu(\boldsymbol{z}) = \psi(\boldsymbol{z}; \boldsymbol{\beta}),$$

donde $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_k)'$ denota a un vector de parámetros desconocidos. Más aún, en muchos casos se supone que $\psi(\cdot; \boldsymbol{\beta})$ es una función lineal de $\boldsymbol{\beta}$ en alguna escala apropiada, *i.e.*

$$\psi(\boldsymbol{z}; \boldsymbol{\beta}) = h(\beta_0 + \beta_1 s_1(\boldsymbol{z}) + ... + \beta_k s_k(\boldsymbol{z}))$$

para alguna transformación uno a uno $h(\cdot)$, conocida, y para algunas funciones suaves, es decir, funciones continuas y derivables en todos su puntos, $\{s_j(\cdot) : j = 1, ..., k\}$, también conocidas. Esta función es tratada entonces como si fuera la verdadera función de regresión $\mu(\cdot)$, por lo que el problema se reduce a hacer inferencias sobre el valor del parámetro $\boldsymbol{\beta}$.

A partir de este planteamiento se obtienen los llamados *modelos lineales generalizados*. El modelo de regresión más usual es aquel donde la variable Y tiene una distribución Normal y $h(\cdot)$ es la función identidad.

El modelo Normal

Supongamos que se tienen n observaciones independientes $(Y_1, \boldsymbol{z}_1), ..., (Y_n, \boldsymbol{z}_n)$ del modelo

$$Y_i \sim N(y_i|\mu(\boldsymbol{z}_i), \sigma^2) \qquad (\sigma^2 > 0, \text{desconocida}),$$

donde

$$\mu(\boldsymbol{z}_i) = \beta_0 + \beta_1 s_1(\boldsymbol{z}_i) + ... + \beta_k s_k(\boldsymbol{z}_i).$$

En particular, el modelo de regresión polinomial más sencillo es de la forma

$$Y_i = \beta_0 + \beta_1 z_i + ... + \beta_k z_i^k + \epsilon_i, \qquad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2) \qquad (i = 1, ..., n)$$

En este caso $z \in \mathbb{R}$ y $s_j(z) = z^j$ para todo $j = 1, .., k$.
Sean $x_{i1} = 1$, $(i = 1, ..., n)$ y

$$x_{ij} = s_j(z_i), \qquad i = 1, .., n; \qquad j = 1, ..., k.$$

Entonces podemos expresar el modelo como

$$Y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} + \epsilon_i, \qquad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2) \qquad (i = 1, ..., n)$$

Resulta conveniente escribir el modelo de forma matricial. Sea $p = k + 1$. Entonces

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$$

donde $\mathbf{Y} = (Y_1, ..., Y_n)'$, $\mathbf{X} = [x_{ij}]$ es una matriz $n \times p$, $\epsilon = (\epsilon_1, ..., \epsilon_n)'$, e $\mathbf{I}_n$ denota a la matriz identidad de orden n. Dicho de otra forma

$$\mathbf{Y} \sim N_n(\mathbf{y}|\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n). \tag{2.5}$$

En el resto del presente trabajo se supondrá que la matriz $\mathbf{X}$ es de rango completo, p.

La función de verosimilitud para el modelo anterior es de la forma

$$L(\boldsymbol{\beta}, \sigma^2; \boldsymbol{y}) \propto (2\pi\sigma^2)^{-n/2} \exp\{-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta})\}.$$

Recordemos que los estimadores de máxima verosimilitud para $\boldsymbol{\beta}$ y σ^2 están dados por $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y}$ y $\widehat{\sigma}^2 = \frac{1}{n}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})$, respectivamente, y notemos que

$$(\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta}) = (\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) + (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})'\boldsymbol{X'X}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}).$$

Así, la función de verosimilitud puede escribirse como

$$L(\boldsymbol{\beta}, \sigma^2; \boldsymbol{y}) \propto (\sigma^2)^{-n/2} \exp\{-\frac{1}{2\sigma^2}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})'\boldsymbol{X'X}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) + n\widehat{\sigma}^2]\}.$$

Para facilitar la notación y el desarrollo subsecuente, es conveniente trabajar en términos de la precisión $\tau = 1/\sigma^2$ en lugar de la varianza σ^2. La verosimilitud toma entonces la forma

$$L(\boldsymbol{\beta}, \tau; \boldsymbol{y}) \propto (\tau)^{n/2} \exp\{-\frac{\tau}{2}[(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})'\boldsymbol{X'X}(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) + n\widehat{\sigma}^2]\}.$$

Distribución inicial

Familia conjugada

Dada la forma de verosimilitud, una familia conjugada particularmente conveniente tiene densidades de la forma

$$p(\boldsymbol{\beta}, \tau) \propto \tau^{n_0/2} \exp\{-\frac{\tau}{2}[(\boldsymbol{\beta} - \boldsymbol{b_0})'\boldsymbol{B_0}(\beta - b_0) + s_0\},$$

donde $n_0, s_0 \in \mathbb{R}, \boldsymbol{b}_0 \in \mathbb{R}^p$ y $\boldsymbol{B}_0$ es una matriz $p \times p$ simétrica y positiva semi-definida.

Notemos que , dado el valor de τ, el *kernel* de la densidad condicional $p(\boldsymbol{\beta}|\tau)$ es proporcional al de la densidad $N_p(\boldsymbol{\beta}|\boldsymbol{b}_0, \tau^{-1}\boldsymbol{B}_0{}^{-1})$, *i.e.*

$$p(\boldsymbol{\beta}|\tau) \propto \tau^{p/2} \exp\{-\frac{\tau}{2}(\boldsymbol{\beta} - \boldsymbol{b}_0)'\boldsymbol{B}_0(\boldsymbol{\beta} - \boldsymbol{b}_0)\},$$

de manera que los factores restantes corresponden a la densidad marginal $p(\tau)$, *i.e.*

$$p(\tau) \propto \tau^{(n_0-p)/2} \exp\{-s_0\tau/2\}.$$

Haciendo $a = n_0 - p + 2$ y $d = s_0$, se tiene entonces que

$$\begin{aligned} p(\boldsymbol{\beta}, \tau) &= p(\boldsymbol{\beta}|\tau)p(\tau) \\ &= N_p(\boldsymbol{\beta}|\boldsymbol{b}_0, \tau^{-1}\boldsymbol{B}_0{}^{-1})Ga(\tau|a/2, d/2) \end{aligned} \tag{2.6}$$

Esta distribución es conocida como *Normal-Gamma* y es propia si $a > 0, d > 0$ y $\boldsymbol{B_0}$ es positiva definida.

Distribución inicial no informativa
En situaciones en las que se desea representar un estado de información inicial vaga acerca de $(\boldsymbol{\beta}, \tau)$, es común utilizar algún tipo de distribución inicial "no informativa". Como ya mencionamos, uno de los métodos más populares para obtener dichas distribuciones es la *Regla de Jeffreys*. Para el modelo de regresión

$$\mathbf{Y} \sim N_n(\mathbf{y}|\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$$

puede demostrarse fácilmente que la distribución inicial de Jeffreys es

$$\pi(\beta, \tau) \propto \tau^{(p-2)/2}$$

Esta distribución es impropia y puede obtenerse a partir de la familia conjugada (2.6) haciendo $a = 0$, $d = 0$ y $\boldsymbol{B}_0 = \mathbf{0}$.

Otra distribución inicial no informativa comúnmente usada en modelos de localización y escala es

$$\pi(\boldsymbol{\beta}, \tau) \propto \tau^{-1}$$

la cual corresponde a la *distribución inicial de referencia* obtenida a partir del método de Bernardo (1979). Al igual que la distribución de Jeffreys, esta distribución es impropia y es un caso límite de la familia conjugada (2.6) cuando $a = -p$, $d = 0$ y $\boldsymbol{B}_0 = \mathbf{0}$.
Comentario. En ambos casos $\boldsymbol{B_0} = 0$ implica que la varianza de la distribución inicial de $\boldsymbol{\beta}$ es infinita, lo que generalmente se interpreta como una forma de

representar un estado de información inicial vaga acerca del valor de $\boldsymbol{\beta}$.

Distribución final

Proposición. La distribución final de $(\boldsymbol{\beta}, \tau)$ para el modelo (2.5) si se utiliza una distribución inicial conjugada de la forma (2.6) es

$$p(\boldsymbol{\beta}, \tau|\boldsymbol{y}) = N_p(\boldsymbol{\beta}|\boldsymbol{b}_1, \tau^{-1}\boldsymbol{B}^{-1})Ga(\tau|a_1/2, d_1/2),$$

donde

$$\begin{aligned} \boldsymbol{b}_1 &= (\boldsymbol{X'X} + \boldsymbol{B_0})^{-1}(\boldsymbol{Xy} + \boldsymbol{B_0 b_0}) \\ \boldsymbol{B}_1 &= \boldsymbol{X'X} + \boldsymbol{B_0} \\ a_1 &= n + a \\ d_1 &= (\boldsymbol{y - Xb_1})'(\boldsymbol{y - Xb_1}) + \boldsymbol{(b_1 - b_0)'B_0(b_1 - b_0)} + d \end{aligned} \tag{2.7}$$

Demostración. Por el Teorema de Bayes,

$$p(\boldsymbol{\beta}, \tau|\boldsymbol{y}) \propto p(\boldsymbol{\beta}, \tau)L(\boldsymbol{\beta}, \tau; \boldsymbol{y}).$$

El resultado es evidente si se nota que

$$\begin{gathered} (\boldsymbol{\beta - \widehat{\beta}})'\boldsymbol{X'X}(\boldsymbol{\beta - \widehat{\beta}}) + (\boldsymbol{\beta - b}_0)'\boldsymbol{B}_0(\boldsymbol{\beta - b}_0) = \\ (\boldsymbol{\beta - b}_1)'\boldsymbol{B}_1(\boldsymbol{\beta - b}_1) + \boldsymbol{\widehat{\beta}X'X\widehat{\beta}} + \boldsymbol{b}_0'\boldsymbol{B_0 b_0} - \boldsymbol{b}_1'\boldsymbol{B_1 b_1} \end{gathered}$$

y

$$\begin{gathered} (\boldsymbol{y - X\widehat{\beta}})'(\boldsymbol{y - X\widehat{\beta}}) + \boldsymbol{\widehat{\beta}X'X\widehat{\beta}} + \boldsymbol{b}_0'\boldsymbol{B_0 b_0} - \boldsymbol{b}_1'\boldsymbol{B_1 b_1} = \\ (\boldsymbol{y - Xb_1})'(\boldsymbol{y - Xb_1}) + \boldsymbol{(b_1 - b_0)'B_0(b_1 - b_0)} \end{gathered}$$

De acuerdo con este resultado, la distribución marginal final de τ es

$$p(\tau|\boldsymbol{y}) = Ga(\tau|a_1/2, d_1/2),$$

lo que implica que la correspondiente distribución final para σ^2 es $IGa(\sigma^2|a_1/2, d_1/2)$.

Por otro lado, si se desea hacer inferencias sobre $\boldsymbol{\beta}$ entonces es necesario calcular su distribución marginal final, dada por

$$\begin{aligned} p(\boldsymbol{\beta}|\boldsymbol{y}) &= \int p(\boldsymbol{\beta}, \tau|\boldsymbol{y})d\tau \\ &= \frac{\Gamma((a_1 + p)/2)}{\Gamma(a_1/2)\pi^{p/2}} \det\left\{\frac{1}{d_1}\boldsymbol{B_1}\right\}^{1/2}\left\{1 + \frac{1}{d_1}\boldsymbol{(\beta - b_1)'B_1(\beta - b_1)}\right\}^{-(a_1+p)/2} \end{aligned}$$

En otras palabras

$$p(\boldsymbol{\beta}|\boldsymbol{y}) = St_p(\boldsymbol{\beta}|\boldsymbol{b}_1, \boldsymbol{T}_1^{-1}, a_1), \tag{2.8}$$

donde $\boldsymbol{T}_1 = \frac{a_1}{d_1}\boldsymbol{B}_1$, de manera que la distribución final de $\boldsymbol{\beta}$ es t de Student con a_1 grados de libertad, parámetro de localización $\boldsymbol{b}_1$, y parámetro de escala $\boldsymbol{T}_1^{-1}$.

Distribución final de referencia
Recordemos que la distribución inicial de referencia, $\pi(\boldsymbol{\beta}, \tau) \propto \tau^{-1}$, corresponde a un caso límite de la familia conjugada (2.6) con $\boldsymbol{B}_0 = \boldsymbol{0}$, $a = -p$ y $d = 0$. Aunque es impropia, ésta da lugar a una distribución final propia siempre y cuando $n > p$. De hecho, en este caso se tiene que

$$\begin{aligned} \boldsymbol{b}_1 &= \widehat{\boldsymbol{\beta}} \\ \boldsymbol{B}_1 &= \boldsymbol{X'X} \\ a_1 &= n - p \\ d_1 &= (n-p)\widetilde{\sigma}^2, \end{aligned}$$

donde $\widetilde{\sigma}^2$ es el estimador insesgado usual para σ^2. Por lo tanto,

$$\pi(\boldsymbol{\beta}, \tau|\boldsymbol{y}) = N_p(\boldsymbol{\beta}|\widehat{\boldsymbol{\beta}}, \tau^{-1}(\boldsymbol{X'X})^{-1})Ga(\tau|(n-p)/2, (n-p)\widetilde{\sigma}^2/2),$$

de donde

$$\pi(\boldsymbol{\beta}, \tau|\boldsymbol{y}) = St_p(\boldsymbol{\beta}|\widehat{\boldsymbol{\beta}}, \widetilde{\sigma}^2(\boldsymbol{X'X})^{-1}, n-p).$$

Inferencia y predicción
Inferencia sobre σ^2
De acuerdo con los resultados de la sección anterior, la distribución final de σ^2 es

$$p(\sigma^2|\boldsymbol{y}) = IGa(\sigma^2|a_1/2, d_1/2).$$

En particular, se tiene que

$$\begin{aligned} E(\sigma^2|\boldsymbol{y}) &= \frac{d_1}{a_1 - 2} \\ Var(\sigma^2|\boldsymbol{y}) &= \frac{2{d_1}^2}{(a_1-2)^2(a_1-4)} \\ Moda(\sigma^2|\boldsymbol{y}) &= \frac{d_1}{a_1 + 2} \end{aligned}$$

Estas cantidades pueden servir de base para hacer inferencias sobre σ^2. También es posible construir intervalos de máxima densidad o simplemente reportar algunos percentiles de la distribución final de σ^2.

Notemos que si se utiliza la distribución de referencia estas expresiones se reducen a

$$E(\sigma^2|\boldsymbol{y}) = \frac{(n-p)\tilde{\sigma}^2}{n-p-2}$$

$$\begin{aligned} Var(\sigma^2|\mathbf{y}) &= \frac{2(n-p)^2\tilde{\sigma}^4}{(n-p-2)^2(n-p-4)} \\ Moda(\sigma^2|\mathbf{y}) &= \frac{(n-p)\tilde{\sigma}^2}{n-p+2} \end{aligned}$$

Inferencia sobre $\boldsymbol{\beta}$
La distribución final de $\boldsymbol{\beta}$ está dada por (2.8). En particular,

$$E(\boldsymbol{\beta}|\boldsymbol{y}) = \boldsymbol{b}_1 \qquad \text{si} \qquad a_1 > 1$$

$$Var(\boldsymbol{\beta}|\boldsymbol{y}) = \frac{a_1}{(a_1-2)}\boldsymbol{T}_1^{-1} \qquad = \frac{d_1}{(a_1-2)}\boldsymbol{B}_1^{-1} \qquad \text{si} \qquad a_1 > 2.$$

Por otro lado, si se utiliza la distribución de referencia entonces

$$E(\boldsymbol{\beta}|\boldsymbol{y}) = \widehat{\boldsymbol{\beta}} \qquad \text{si} \qquad n > p+1$$

$$Var(\boldsymbol{\beta}|\boldsymbol{y}) = \frac{(n-p)\tilde{\sigma}^2}{(n-p-2)}(\boldsymbol{X'X})^{-1} \qquad \text{si} \qquad n > p+2$$

Como en el caso anterior, estas cantidades pueden servir de base para hacer inferencias sobre $\boldsymbol{\beta}$. Notemos, sin embargo, que en este caso el interés se centra generalmente en combinaciones lineales de las entradas del vector $\boldsymbol{\beta}$.

Sea $\gamma = \boldsymbol{C\beta}$, donde $\boldsymbol{C}$ es una matriz $r \times p$ de rango $r (r \leq p)$. Entonces

$$p(\boldsymbol{\gamma}|\sigma^2, \boldsymbol{y}) = N_r(\boldsymbol{\gamma}|\boldsymbol{g}, \sigma^2\boldsymbol{G}),$$

donde $\boldsymbol{g} = \boldsymbol{Cb}_1$ y $\boldsymbol{G} = \boldsymbol{CB}_1^{-1}\boldsymbol{C}'$. Por lo tanto

$$p(\boldsymbol{\gamma}|\boldsymbol{y}) = St_t(\boldsymbol{\gamma}|\boldsymbol{g}, (d_1/a_1)\boldsymbol{G}, a_1)$$

Supongamos, por ejemplo, que $r = 1$ y $\boldsymbol{C} = \boldsymbol{e}_i' = (0, ..., 1, ..., 0)$ para alguna $i = 1, ..., p$. Entonces $\boldsymbol{\gamma} = \boldsymbol{e}_i'\boldsymbol{\beta} = \beta_{i-1}$. En este caso $g = \boldsymbol{e}_i'\boldsymbol{b_1} = b_{1,i-1}$ y $G = B_1^{ii}$, donde B_1^{ii} es la entrada (i,i) de la matriz $\boldsymbol{B}_1^{-1}$. Por lo tanto,

$$p(\beta_j|\boldsymbol{y}) = St(\beta_j|b_{1j}, (d_1/a_1)B_1^{j+1,j+1}, a_1) \qquad (j = 0, 1, ..., k)$$

Predicción
Supongamos que se desea predecir Y_*, un nuevo valor de la variable de respuesta, dado el vector de covariables $\boldsymbol{x}_*' = (1, x_{1*}, ..., x_{k*})$. De acuerdo con el modelo,

$$Y_* = \beta_0 + \beta_1 x_{1*} + ... + \beta_k x_{k*} + \epsilon_* = \boldsymbol{x'*\beta} + \epsilon*,$$

donde $\epsilon_* \sim N(0, \sigma^2)$ es independiente de $\boldsymbol{\epsilon}$. Entonces,

$$\mu_* \overset{\{def\}}{=} E(Y_*|\boldsymbol{\beta}, \boldsymbol{\sigma}^2) = \boldsymbol{x}_*'\boldsymbol{\beta}$$

y

$$Var(Y_*|\boldsymbol{\beta}, \boldsymbol{\sigma}^2) = \boldsymbol{\sigma}^2$$

El problema de predicción puede abordarse de dos maneras:
(a) *Inferencia sobre* μ_*. El parámetro μ_*, que corresponde al valor esperado de la observación futura Y_*, no es más que una combinación lineal de los coeficientes de regresión.

Sea $r = 1$ y $\mathbf{C} = \mathbf{x}'_*$. Entonces $\gamma = \mathbf{x}'_*\boldsymbol{\beta} = \mu*$, $g = \mathbf{x}'_*\mathbf{b}_1$ y $G = \mathbf{x}'_*\mathbf{B}_1^{-1}\mathbf{x}_*$, por lo que

$$p(\mu_*|\mathbf{y}) = St(\mu_*|\mathbf{x}'_*\mathbf{b}_1, (d_1/a_1)\mathbf{x}'_*\mathbf{B}_1^{-1}\mathbf{x}_*, a_1)$$

En particular, si se utiliza la distribución de referencia entonces

$$\pi(\mu_*|\mathbf{y}) = St(\mu_*|\mathbf{x}'_*\widehat{\boldsymbol{\beta}}, \widetilde{\sigma}^2\mathbf{x}'_*(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_*, n-p)$$

En este caso, el intervalo de máxima densidad del $(1-\alpha)\times 100\,\%$ está dado por

$$\mathbf{x}'_*\widehat{\boldsymbol{\beta}} \pm t_{(n-p)}^{1-\alpha/2}\widetilde{\sigma}^2\sqrt{\mathbf{x}'_*(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_*},$$

donde $t_{(n-p)}^{1-\alpha/2}$ es el cuantil de orden $(1-\alpha/2)$ de una distribución t de Student estandarizada con $(n-p)$ grados de libertad. Este intervalo tiene la misma forma que el correspondiente intervalo frecuentista.
(a) *Inferencia sobre* Y_*. En este caso interesa calcular la distribución predictiva final para Y_*, *i.e.*

$$p(y_*|\mathbf{y}) = \int\int p(y_*|\boldsymbol{\beta}, \sigma^2)p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y})d\boldsymbol{\beta}d\sigma^2.$$

Recordemos primero que $Y_* = \boldsymbol{x}'_*\boldsymbol{\beta} + \epsilon_*$. Trabajando condicionalmente en σ^2, tenemos que $\boldsymbol{\beta}$ y ϵ_* son independientes y

$$p(\boldsymbol{\beta}|\sigma^2, \boldsymbol{y}) = N_p(\boldsymbol{\beta}|\boldsymbol{b}_1, \sigma^2\boldsymbol{B}_1^{-1})$$

$$p(\epsilon_*|\sigma^2, \boldsymbol{y}) = N(\epsilon_*|0, \sigma^2).$$

Esto implica que

$$p(y_*|\sigma^2, \boldsymbol{y}) = N(y_*|\boldsymbol{x}'_*\boldsymbol{b}_1, \sigma^2 1 + \boldsymbol{x}'_*\boldsymbol{B}_1^{-1}\boldsymbol{x}_*, a_1).$$

Finalmente, integrando con respecto a la distribución final de σ^2,

$$p(y_*|\mathbf{y}) = St(y_*|\mathbf{x}'_*\mathbf{b}_1, (d_1/a_1)\{1 + \mathbf{x}'_*\mathbf{B}_1^{-1}\mathbf{x}_*\}, a_1).$$

Si se utiliza la distribución de referencia entonces la distribución predictiva final toma la forma

$$p(y_*|\boldsymbol{y}) = St(y_*|\boldsymbol{x_*}'\boldsymbol{\beta}, \sigma^2(1 + \boldsymbol{x}_*'\{\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_*\}, n-p).$$

En este caso, el intervalo de máxima densidad del $(1-\alpha)\times 100\,\%$ está dado por

$$\boldsymbol{x}_*'\widehat{\boldsymbol{\beta}} \pm t_{(n-p)}^{1-\alpha/2}\widetilde{\sigma}^2\sqrt{1 + \mathbf{x}_*'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_*}.$$

Como en el caso anterior, este intervalo tiene la misma forma que el correspondiente intervalo frecuentista.

2.1.3. Regresiones aparentemente no relacionadas

El modelo de regresiones aparentemente no relacionadas propuesto por Zellner, puede ser visto como un caso especial de los modelos de regresión generalizados $E(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{\beta}$, $Var(\boldsymbol{y}) = \sigma^2\boldsymbol{\Omega}$, donde la matriz $\boldsymbol{\Omega}$ generalmente contiene parámetros desconocidos que deben ser estimados; el estimador más común de la matriz de covarianza es el estimador de mínimos cuadrados $\hat{\boldsymbol{\beta}}_{MC}$.

El modelo SUR básico supone que para cada observación i existen M variables dependientes $y_{i1}, ...y_{ij}, ...y_{iM}$ disponibles, cada una con su propio modelo de regresión lineal:

$$y_{ij} = \boldsymbol{x}_{ij}'\boldsymbol{\beta}_j + \epsilon_{ij} \qquad i = 1, ..., N$$

o si escribimos las N observaciones de cada variable dependiente como un vector,

$$\boldsymbol{y}_j = \boldsymbol{X}_j\boldsymbol{\beta}_j + \boldsymbol{\epsilon}_j$$

para $j = 1, ..., M$, donde $\boldsymbol{y}_j$ y $\boldsymbol{\epsilon}_j$ son vectores de dimensión N y $\boldsymbol{X}_j$ es una matriz $N \times K_j$, donde K_j =dim(β_j) es el número de regresores para la regresión j-ésima.

De esta manera, tenemos el siguiente modelo de regresión multivariada:

$$[\boldsymbol{y}_1 \vdots \cdots \vdots \boldsymbol{y}_M]_{(N\times M)}$$
$$= [\boldsymbol{X}_1 \vdots \cdots \vdots \boldsymbol{X}_M]_{(N\times K)} \begin{pmatrix} \boldsymbol{\beta}_1 & 0 & 0 \\ \vdots & \ddots & \vdots \\ 0 & 0 & \boldsymbol{\beta}_M \end{pmatrix}_{(K\times M)} + [\boldsymbol{\epsilon}_1 \vdots \cdots \vdots \boldsymbol{\epsilon}_M]_{(N\times M)}$$

Las condiciones estándar para el modelo clásico de regresión suponen para cada j:

$$E(\boldsymbol{y}_j) = \boldsymbol{X}_j\boldsymbol{\beta}_j,$$
$$Var(\boldsymbol{y}_j) = \sigma_{jj}\boldsymbol{I}_N,$$

con $\boldsymbol{X}_j$ no estocástica y $rango(\boldsymbol{X}_j) = K_j$. Bajo estas condiciones, y la condición adicional de multinormalidad de $\boldsymbol{y}_j$, la teoría de inferencia usual es válida para el estimador de mínimos cuadrados de $\boldsymbol{\beta}_j$, aplicada a cada ecuación por separado.

Sin embargo, el modelo SUR permite una covarianza diferente de cero entre los errores ϵ_{ij} y ϵ_{ik} para cada individuo i, a lo largo de las ecuaciones j y k, es decir,

$$Cov(\epsilon_{ij}, \epsilon_{ik}) = \sigma_{ij}$$

suponiendo

$$Cov(\epsilon_{ij}, \epsilon_{i'k}) = 0$$

si $i \neq i'$. Esto puede ser expresado de forma compacta como

$$Cov(\boldsymbol{\epsilon}_j, \boldsymbol{\epsilon}_k) = \sigma_{jk}\boldsymbol{I}_N.$$

Producto Kronecker

La idea de Zellner fue que, dada la forma usual de acomodar las variables dependientes y_{ij} dentro de un vector $\boldsymbol{y}_j$ de dimensión N, estos últimos vectores a su vez pueden ser acomodados en una vector $\boldsymbol{y}$ de dimensión NM con su arreglo correspondiente para los errores, los coeficientes del vector y los regresores:

$$\boldsymbol{y}_{(MN\times 1)} = \begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \dots \\ \boldsymbol{y}_M \end{pmatrix}, \epsilon_{(MN\times 1)} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \dots \\ \epsilon_M \end{pmatrix}, \boldsymbol{\beta}_{(K\times 1)} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \dots \\ \boldsymbol{\beta}_M \end{pmatrix},$$

y

$$\boldsymbol{X}_{(MN\times K)} = \begin{pmatrix} \boldsymbol{X}_1 & 0 & \dots & 0 \\ 0 & \boldsymbol{X}_1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & \boldsymbol{X}_M \end{pmatrix},$$

con $K = \sum_{j=1}^{M} K_j$.

Con esta notación y los supuestos para cada ecuación j se sigue que:

$$E(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{\beta}$$

y

$$Var(\boldsymbol{y})_{(MN\times MN)} = \begin{pmatrix} \sigma_{11}\boldsymbol{I}_N & \sigma_{12}\boldsymbol{I}_N & \dots & \sigma_{1M}\boldsymbol{I}_N \\ \sigma_{21}\boldsymbol{I}_N & \sigma_{22}\boldsymbol{I}_N & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \sigma_{M1}\boldsymbol{I}_N & \dots & \dots & \sigma_{MM}\boldsymbol{I}_N \end{pmatrix}.$$

Esta matriz de covarianza es una combinación particular de la matriz

$$\Sigma_{(M\times M)} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1M} \\ \sigma_{21} & \sigma_{22} & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \sigma_{M1} & \dots & \dots & \sigma_{MM} \end{pmatrix}.$$

y la matriz identidad $\boldsymbol{I}_N$. Un sistema de notación para tales combinaciones fue propuesto por el matemático Kronecker, de donde viene el nombre de *Producto Kronecker*; para dos matrices $\boldsymbol{A} \equiv [a_{ij}]$ $(i = 1, ..., L, j = 1, ..., M)$ y $\boldsymbol{B}$, el producto Kronecker de $\boldsymbol{A}$ y $\boldsymbol{B}$ está definido como

$$\boldsymbol{A} \otimes \boldsymbol{B} \equiv \begin{pmatrix} a_{11}\boldsymbol{B} & a_{12}\boldsymbol{B} & \dots & a_{1M}\boldsymbol{B} \\ a_{21}\boldsymbol{B} & a_{22}\boldsymbol{B} & \dots & \dots \\ \dots & \dots & \dots & \dots \\ a_{L1}\boldsymbol{B} & \dots & \dots & a_{LM}\boldsymbol{B} \end{pmatrix},$$

Con esta notación, claramente

$$Var(\boldsymbol{y}) = \Sigma \otimes \boldsymbol{I}_N$$

para el modelo SUR.

El producto Kronecker satisface una regla distributiva, en la cual:

$$(\boldsymbol{A} \otimes \boldsymbol{B})(\boldsymbol{C} \otimes \boldsymbol{D}) = \boldsymbol{AC} \otimes \boldsymbol{BD},$$

suponiendo que todos los productos de matrices están bien definidos. De esta regla se sigue que para productos Kronecker inversos:

$$(\boldsymbol{A} \otimes \boldsymbol{B})^{-1} = \boldsymbol{A}^{-1} \otimes \boldsymbol{B}^{-1}$$

suponiendo $\boldsymbol{A}$ y $\boldsymbol{B}$ invertibles.

Mínimos cuadrados

Con la notación anterior, el estimador clásico de mínimos cuadrados para el vector $\boldsymbol{\beta}$ puede ser expresado como:

$$\hat{\boldsymbol{\beta}}_{MC} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y} = \begin{pmatrix} (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{y}_1 \\ (\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{y}_2 \\ \dots \\ (\boldsymbol{X}_M'\boldsymbol{X}_M)^{-1}\boldsymbol{X}_M'\boldsymbol{y}_M \end{pmatrix},$$

En contraste, el estimador generalizado de mínimos cuadrados para $\boldsymbol{\beta}$ (asumiendo Σ conocido) es:

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{MCG} &= (\boldsymbol{X'}(\textstyle\sum \otimes \boldsymbol{I}_N)^{-1}\boldsymbol{X})^{-1}\boldsymbol{X'}(\textstyle\sum \otimes \boldsymbol{I_N})^{-1}\boldsymbol{y} \\ &= (\boldsymbol{X'}(\textstyle\sum^{-1} \otimes \boldsymbol{I_N})\boldsymbol{X})^{-1}\boldsymbol{X'}(\textstyle\sum^{-1} \otimes \boldsymbol{I}_N)\boldsymbol{y} \\ &= \begin{pmatrix} \sigma^{11}(\boldsymbol{X}_1'\boldsymbol{X}_1) & \sigma^{12}(\boldsymbol{X}_1'\boldsymbol{X}_2) & \dots & \sigma^{1M}(\boldsymbol{X}_1'\boldsymbol{X}_M) \\ \sigma^{21}(\boldsymbol{X}_2'\boldsymbol{X}_1) & \sigma^{22}(\boldsymbol{X}_2'\boldsymbol{X}_2) & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \sigma^{M1}(\boldsymbol{X}_M'\boldsymbol{X}_1) & \dots & \dots & \sigma^{MM}(\boldsymbol{X}_M'\boldsymbol{X}_M) \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{X}_1'(\sum_j \sigma^{1j}\boldsymbol{y}_j) \\ \boldsymbol{X}_2'(\sum_j \sigma^{2j}\boldsymbol{y}_j) \\ \dots \\ \boldsymbol{X}_M'(\sum_j \sigma^{Mj}\boldsymbol{y}_j) \end{pmatrix} \end{aligned}$$

donde σ^{ij} está definido como el elemento en la i-ésima fila y la j-ésima columna de $\mathbf{\Sigma}^{-1}$, es decir, $\mathbf{\Sigma}^{-1} \equiv [\sigma^{ij}]$.

Para tener una mejor idea del estimador de mínimos cuadrados generalizado considere el caso especial $M = 2$ con

$$\hat{\boldsymbol{\beta}}_{MC} \equiv \begin{pmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \end{pmatrix}, \hat{\boldsymbol{\beta}}_{MCG} \equiv \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix};$$

podemos mostrar que los estimadores por mínimos cuadrados generalizados $\boldsymbol{\beta}_1$ y $\boldsymbol{\beta}_2$ satisfacen las dos ecuaciones:

$$\boldsymbol{\beta}_1 = \boldsymbol{b}_1 - \Big(\frac{\sigma_{21}}{\sigma_{22}}\Big)(\boldsymbol{X}_1'\boldsymbol{X}_1)\boldsymbol{X}_1'(\boldsymbol{y}_2 - \boldsymbol{X}_2'\hat{\boldsymbol{\beta}}_2)$$

$$\boldsymbol{\beta}_2 = \boldsymbol{b}_2 - \Big(\frac{\sigma_{12}}{\sigma_{11}}\Big)(\boldsymbol{X}_2'\boldsymbol{X}_2)\boldsymbol{X}_2'(\boldsymbol{y}_1 - \boldsymbol{X}_1'\hat{\boldsymbol{\beta}}_1).$$

Por lo que los estimadores de mínimos cuadrados generalizados pueden ser vistos como versiones ajustadas del estimador clásico de mínimos cuadrados, donde el ajuste contiene la regresión de los residuales de mínimos cuadrados generalizados de la otra ecuación en los regresores de cada ecuación.

Un caso importante es cuando la matriz Σ es diagonal, es decir, $\sigma_{ij} = 0$ si $i \neq j$. En este caso, si Σ^{-1} =diag$[1/\sigma_{ii}]$, se sigue que

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{MCG} &= \begin{pmatrix} \frac{1}{\sigma_{11}}(\boldsymbol{X}_1'\boldsymbol{X}_1) & 0 & \dots & 0 \\ 0 & \frac{1}{\sigma_{22}}(\boldsymbol{X}_2'\boldsymbol{X}_2) & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \frac{1}{\sigma_{MM}}(\boldsymbol{X}_M'\boldsymbol{X}_M) \end{pmatrix}^{-1} \begin{pmatrix} \frac{1}{\sigma_{11}}(\boldsymbol{X}_1'\boldsymbol{y}_1) \\ \frac{1}{\sigma_{22}}(\boldsymbol{X}_2'\boldsymbol{y}_2) \\ \dots \\ \frac{1}{\sigma_{MM}}(\boldsymbol{X}_M'\boldsymbol{y}_M) \end{pmatrix} \\
&= \begin{pmatrix} (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{y}_1 \\ (\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'\boldsymbol{y}_2 \\ \dots \\ (\boldsymbol{X}_M'\boldsymbol{X}_M)^{-1}\boldsymbol{X}_M'\boldsymbol{y}_M \end{pmatrix} \\
&= \hat{\boldsymbol{\beta}}_{MC} \qquad (2.9)
\end{aligned}$$

Otro caso importante es cuando la matriz de regresores es idéntica para cada ecuación, es decir, $\boldsymbol{X}_j \equiv \boldsymbol{X}_0$ para alguna matriz $\boldsymbol{X}_0$ de $N \times K*$ con $K* = K/M$. Aquí la matriz $\boldsymbol{X}$ toma la forma:

$$\begin{aligned}
\boldsymbol{X}_{(MN \times K)} &= \begin{pmatrix} \boldsymbol{X}_0 & 0 & \dots & 0 \\ 0 & \boldsymbol{X}_0 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \boldsymbol{X}_0 \end{pmatrix} \\
&= (\boldsymbol{I}_M \otimes \boldsymbol{X}_0),
\end{aligned}$$

y el estimador de mínimos cuadrados generalizados también se reduce al clásico de mínimos cuadrados:

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{MCG} &= (\boldsymbol{X}'(\Sigma \otimes \boldsymbol{I}_N)^{-1}\boldsymbol{X}'(\Sigma \otimes \boldsymbol{I}_N)^{-1}\boldsymbol{y}) \\
&= ((\boldsymbol{I}_M \otimes \boldsymbol{X})'(\Sigma^{-1} \otimes \boldsymbol{I}_N)(\boldsymbol{I}_M \otimes \boldsymbol{X}))^{-1}(\boldsymbol{I}_M \otimes \boldsymbol{X})'(\Sigma^{-1} \otimes \boldsymbol{I}_N)\boldsymbol{y} \\
&= (\Sigma^{-1} \otimes \boldsymbol{X}'\boldsymbol{X})^{-1}(\Sigma^{-1} \otimes \boldsymbol{X}')\boldsymbol{y} \\
&= (\boldsymbol{I}_M \otimes (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}') \begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \cdots \\ \boldsymbol{y}_M \end{pmatrix} \\
&= \begin{pmatrix} (\boldsymbol{X}_0'\boldsymbol{X}_0)^{-1}\boldsymbol{X}_0'\boldsymbol{y}_1) \\ (\boldsymbol{X}_0'\boldsymbol{X}_0)^{-1}\boldsymbol{X}_0'\boldsymbol{y}_2) \\ \cdots \\ (\boldsymbol{X}_0'\boldsymbol{X}_0)^{-1}\boldsymbol{X}_0'\boldsymbol{y}_M) \end{pmatrix} \\
&= \hat{\boldsymbol{\beta}}_{MC}
\end{aligned}$$

2.1.4. Métodos computacionales

La implementación de las técnicas Bayesianas usualmente requiere de un esfuerzo computacional muy alto. La mayor parte de este esfuerzo se concentra en el cálculo de ciertas características de la distribución final del parámetro de interés. Así, por ejemplo, para pasar de una distribución conjunta a una colección de distribuciones y momentos marginales que sean útiles para hacer inferencias sobre subconjuntos de parámetros, se requiere integrar. Sin embargo, en la práctica es común que la dimensión de θ sea muy grande. Por otro lado, excepto en aplicaciones muy sencillas, tanto $p(x|\theta)$ como $p(\theta)$ pueden llegar a tener formas muy complicadas. En la gran mayoría de los problemas, las integrales requeridas no pueden resolverse analíticamente, por lo que es necesario contar con métodos numéricos eficientes que permitan calcular o aproximar integrales en varias dimensiones. Entre los métodos numéricos, los que destacan son: aproximación de Laplace, cuadratura (integración numérica), métodos Monte Carlo y Monte Carlo vía cadenas de Markov.

En este trabajo hablaré solamente sobre los métodos Monte Carlo y Monte Carlo vía cadenas de Markov.

Integración Monte Carlo

Considere la integral

$$I = \int_B f(x)dx = \int_B \left\{\frac{f(x)}{\pi(x)}\right\}\pi(x)dx = \int_B g(x)\pi(x)dx = E[g(x)],$$

la idea básica del método Monte Carlo consiste en escribir la integral requerida como el valor esperado de alguna función $g(\cdot)$ con respecto a alguna distribución de probabilidad $\pi(\cdot)$, es decir, evalúa $E[g(X)]$ obteniendo muestras $X_t, t = 1, ..., n$ de $\pi(\cdot)$ y después aproxima

$$E[g(X)] \approx \frac{1}{n} \sum_{t=1}^{n} g(X_t)$$

Por lo tanto, la media poblacional de $g(X)$ es estimada por una media muestral. Cuando las muestras X_t son independientes, la ley de los grandes números asegura que la aproximación puede ser tan exacta como se quiera incrementando el tamaño de muestra n.

Las X_t se pueden generar por cualquier proceso en el cual se den muestras a través del soporte de $\pi(\cdot)$ en las proporciones correctas. Una manera de hacer esto es por cadenas de Markov teniendo $\pi(\cdot)$ como una distribución estacional; y esto es lo que se conoce como Monte Carlo vía Cadenas de Markov (MCMC).

Monte Carlo vía cadenas de Markov

Una cadena de Markov es un tipo especial de proceso estocástico, y como tal se refiere a las características de sucesiones de variables aleatorias. Un proceso estocástico puede ser definido como una colección de cantidades aleatorias $\{\theta^{(t)} : t \in T\}$ para algún conjunto de T. El conjunto $\{\theta^{(t)} : t \in T\}$ es conocido como un proceso estocástico con espacio de estados S y conjunto de índices o parámetros T.

En términos simples, una Cadena de Markov es un proceso estocástico donde dado el presente estado, los estados pasado y futuro son independientes. Es decir, una Cadena de Markov es una sucesión de variables aleatorias $X_0, X_1, X_2, ...$ tal que, en cada tiempo $t \geq 0$, el siguiente estado X_{t+1} es muestreado de una distribución $p(X_{t+1}|X_t)$, la cual solo depende del estado actual X_t. Esto es, dado X_t, el siguiente estado X_{t+1} no depende de la historia de la cadena $X_0, ..., X_{t-1}$. Se asume también que la cadena es homogénea en el tiempo: esto es, $p(\cdot|\cdot)$ no depende de t.

Ahora bien, las técnicas de Monte Carlo vía cadenas de Markov permiten generar, de manera iterativa, observaciones de distribuciones multivariadas que difícilmente podrían simularse utilizando métodos directos. La idea básica es muy simple: construir una cadena de Markov que sea fácil de simular y cuya distribución de equilibrio corresponda a la distribución final que nos interesa. Entre los métodos más discutidos para la construcción de dichas cadenas están el algoritmo de Metropolis-Hastings y el muestreo de Gibbs.

Proposición 2.1 *Sea* $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots$ *una cadena de Markov* homogénea, irreducible y aperiódica*, con espacio de estados* Θ *y distribución de equilibrio*

$p(\boldsymbol{\theta}|\boldsymbol{x})$. *Entonces, conforme* $t \to \infty$,

(*i*) $\boldsymbol{\theta}^{(t)} \xrightarrow{\mathcal{D}} \boldsymbol{\theta}$, *donde* $\boldsymbol{\theta} \sim p(\boldsymbol{\theta}|\boldsymbol{x})$;

(*ii*) $\frac{1}{t}\sum_{i=1}^{t} g(\boldsymbol{\theta}^{(i)}) \to E(g(\boldsymbol{\theta})|\boldsymbol{x})$.

Algoritmo de Metropolis-Hastings

Este algoritmo construye una cadena de Markov apropiada definiendo las probabilidades de transición de la siguiente manera.

Sea $Q(\boldsymbol{\theta}^*|\boldsymbol{\theta})$ una distribución de transición (arbitraria) y definamos

$$\alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = \text{mín}\left\{\frac{p(\boldsymbol{\theta}^*|\boldsymbol{x})\, Q(\boldsymbol{\theta}|\boldsymbol{\theta}^*)}{p(\boldsymbol{\theta}|\boldsymbol{x})\, Q(\boldsymbol{\theta}^*|\boldsymbol{\theta})}, 1\right\}.$$

Algoritmo. Dado un valor inicial $\boldsymbol{\theta}^{(0)}$, la t-ésima iteración consiste en:

1. generar una observación $\boldsymbol{\theta}^*$ de $Q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t)})$;
2. generar una variable $u \sim U(0,1)$;
3. si $u \leq \alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta}^{(t)})$, hacer $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^*$; en caso contrario, hacer $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)}$.

Este procedimiento genera una cadena de Markov con distribución de transición

$$\begin{aligned} P(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)}) &= \alpha(\boldsymbol{\theta}^{(t+1)}, \boldsymbol{\theta}^{(t)})\, Q(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)}) \\ &+ I\{\theta^{(t+1)} = \theta^{(t)}\} \int \alpha(\theta^*|\theta^{(t)}) Q(\theta^*|\theta^{(t)}) d\theta^*. \end{aligned}$$

La probabilidad de aceptación $\alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta})$ sólo depende de $p(\boldsymbol{\theta}|\boldsymbol{x})$ a través de un cociente, por lo que la constante de normalización no es necesaria.

Comentario. La versión original del algoritmo de Metropolis toma $Q(\boldsymbol{\theta}^*|\boldsymbol{\theta}) = Q(\boldsymbol{\theta}|\boldsymbol{\theta}^*)$, en cuyo caso

$$\alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = \text{mín}\left\{\frac{p(\boldsymbol{\theta}^*|\boldsymbol{x})}{p(\boldsymbol{\theta}|\boldsymbol{x})}, 1\right\}.$$

Dos casos particulares utilizados comúnmente en la práctica son:

$\diamond$ *Caminata aleatoria.* Sea $Q(\boldsymbol{\theta}^*|\boldsymbol{\theta}) = Q_1(\boldsymbol{\theta}^* - \boldsymbol{\theta})$, donde $Q_1(\cdot)$ es una densidad de probabilidad simétrica centrada en el origen. Entonces

$$\alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = \text{mín}\left\{\frac{p(\boldsymbol{\theta}^*|\boldsymbol{x})}{p(\boldsymbol{\theta}|\boldsymbol{x})}, 1\right\}.$$

$\diamond$ *Independencia.* Sea $Q(\boldsymbol{\theta}^*|\boldsymbol{\theta}) = Q_0(\boldsymbol{\theta}^*)$, donde $Q_0(\cdot)$ es una densidad de probabilidad sobre Θ. Entonces

$$\alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta}) = \text{mín}\left\{\frac{\omega(\boldsymbol{\theta}^*)}{\omega(\boldsymbol{\theta})}, 1\right\},$$

con $\omega(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\boldsymbol{x})/Q_0(\boldsymbol{\theta})$.

En la práctica es común utilizar, después de una reparametrización apropiada, distribuciones de transición normales ó t de Student ligeramente sobredispersas, *e.g.*

$$Q(\boldsymbol{\theta}^*|\boldsymbol{\theta}) = N_d(\boldsymbol{\theta}^*|\boldsymbol{\theta}, \kappa\, \boldsymbol{V}(\hat{\boldsymbol{\theta}})) \qquad \text{(caminata aleatoria)}$$

ó

$$Q_0(\boldsymbol{\theta}^*) = N_d(\boldsymbol{\theta}^*|\hat{\boldsymbol{\theta}}, \kappa\, \boldsymbol{V}(\hat{\boldsymbol{\theta}})) \qquad \text{(independencia),}$$

donde $\hat{\boldsymbol{\theta}}$ y $\boldsymbol{V}(\hat{\boldsymbol{\theta}})$ denotan a la media y a la matriz de varianzas-covarianzas de la aproximación normal asintótica para $p(\boldsymbol{\theta}|\boldsymbol{x})$, respectivamente, y $\kappa \geq 1$ es un factor de sobredispersión.

Muestreo Gibbs

Al igual que el algoritmo de Metropolis, el algoritmo de Gibbs permite simular una cadena de Markov $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots$ con distribución de equilibrio $p(\boldsymbol{\theta}|\boldsymbol{x})$. En este caso, sin embargo, cada nuevo valor de la cadena se obtiene a través de un proceso iterativo que sólo requiere generar muestras de distribuciones cuya dimensión es menor que d y que en la mayoría de los casos tienen una forma más sencilla que la de $p(\boldsymbol{\theta}|\boldsymbol{x})$.

Sea $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k)$ una partición del vector $\boldsymbol{\theta}$, donde $\boldsymbol{\theta}_i \in \mathbb{R}^{d_i}$ y $\sum_{i=1}^{k} d_i = d$. Las densidades

$$\begin{aligned}
&p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_k, \boldsymbol{x}) \\
&\qquad\vdots \\
&p(\boldsymbol{\theta}_i|\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{i-1}, \boldsymbol{\theta}_{i+1}, \ldots, \boldsymbol{\theta}_k, \boldsymbol{x}) \qquad (i = 2, \ldots, k-1) \\
&\qquad\vdots \\
&p(\boldsymbol{\theta}_k|\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{k-1}, \boldsymbol{x})
\end{aligned}$$

se conocen como *densidades condicionales completas* y en general pueden identificarse fácilmente al inspeccionar la forma de la distribución final $p(\boldsymbol{\theta}|\boldsymbol{x})$. De hecho, para cada $i = 1, \ldots, k$,

$$p(\boldsymbol{\theta}_i|\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{i-1}, \boldsymbol{\theta}_{i+1}, \ldots, \boldsymbol{\theta}_k, \boldsymbol{x}) \propto p(\boldsymbol{\theta}|\boldsymbol{x}),$$

donde $p(\boldsymbol{\theta}|\boldsymbol{x}) = p(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k|\boldsymbol{x})$ es vista sólo como función de $\boldsymbol{\theta}_i$.

Dado un valor inicial $\boldsymbol{\theta}^{(0)} = (\boldsymbol{\theta}_1^{(0)}, \ldots, \boldsymbol{\theta}_k^{(0)})$, el algoritmo de Gibbs simula una cadena de Markov en la que $\boldsymbol{\theta}^{(t+1)}$ se obtiene a partir de $\boldsymbol{\theta}^{(t)}$ de la siguiente manera:

generar una observación $\boldsymbol{\theta}_1^{(t+1)}$ de $p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2^{(t)}, \boldsymbol{\theta}_3^{(t)}, \ldots, \boldsymbol{\theta}_k^{(t)}, \boldsymbol{x})$;

generar una observación $\boldsymbol{\theta}_2^{(t+1)}$ de $p(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\theta}_3^{(t)}, \ldots, \boldsymbol{\theta}_k^{(t)}, \boldsymbol{x})$;

$\vdots$

generar una observación $\boldsymbol{\theta}_k^{(t+1)}$ de $p(\boldsymbol{\theta}_k|\boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\theta}_2^{(t+1)}, \ldots, \boldsymbol{\theta}_{k-1}^{(t+1)}, \boldsymbol{x})$.

La sucesión $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots$ así obtenida es entonces una realización de una cadena de Markov cuya distribución de transición está dada por

$$P(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)}) = \prod_{i=1}^{k} p(\boldsymbol{\theta}_i^{(t+1)}|\boldsymbol{\theta}_1^{(t+1)}, \ldots, \boldsymbol{\theta}_{i-1}^{(t+1)}, \boldsymbol{\theta}_{i+1}^{(t)}, \ldots, \boldsymbol{\theta}_k^{(t)}, \boldsymbol{x}).$$

Comentario. En ocasiones la distribución final implica cierta estructura de independencia condicional entre algunos de los elementos del vector $\boldsymbol{\theta}$. Es estos casos es común que muchas de las densidades condicionales completas se simplifiquen.

Ejemplo 2.1 Consideremos el modelo jerárquico definido por

$$\begin{array}{ll} \text{I.} & p(\boldsymbol{x}|\boldsymbol{\omega}) = \prod_{i=1}^{m} p(x_i|\omega_i); \\ \text{II.} & p(\boldsymbol{\omega}|\phi) = \prod_{i=1}^{m} p(\omega_i|\phi); \\ \text{III.} & p_0(\phi). \end{array}$$

Esta estructura define un modelo para $\boldsymbol{x}$ parametrizado por $\boldsymbol{\theta} = (\boldsymbol{\omega}, \phi) = (\omega_1, \ldots, \omega_m, \phi)$ y con distribución inicial $p(\boldsymbol{\theta}) = p_0(\phi)\, p(\boldsymbol{\omega}|\phi)$, de manera que la distribución final está dada por

$$p(\boldsymbol{\theta}|\boldsymbol{x}) \propto p_0(\phi) \prod_{i=1}^{m} \{p(x_i|\omega_i)\, p(\omega_i|\phi)\}.$$

Entonces $k = m+1$ y las densidades condicionales completas toman la forma

$$\begin{array}{rcl} p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_{k-1}, \boldsymbol{\theta}_k, \boldsymbol{x}) & = & p(\omega_1|\phi, x_1) \\ \vdots & & \vdots \\ p(\boldsymbol{\theta}_{k-1}|\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{k-2}, \boldsymbol{\theta}_k, \boldsymbol{x}) & = & p(\omega_m|\phi, x_m) \\ p(\boldsymbol{\theta}_k|\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{k-2}, \boldsymbol{\theta}_{k-1}, \boldsymbol{x}) & = & p_0(\phi|\boldsymbol{\omega}), \end{array}$$

donde $p_0(\phi|\boldsymbol{\omega}) \propto p_0(\phi)\, p(\boldsymbol{\omega}|\phi)$. ◁

El siguiente ejemplo presenta una aplicación interesante del muestreo Gibbs al problema de observaciones faltantes.

Ejemplo 2.2 Sea $\boldsymbol{x} = (\boldsymbol{z}, \tilde{\boldsymbol{z}})$ un conjunto de observaciones del modelo

$$\{p(x|\boldsymbol{\theta}), p(\boldsymbol{\theta})\},$$

y supongamos que $\boldsymbol{z}$ denota a los *datos observados*, mientras que $\tilde{\boldsymbol{z}}$ representa a las *observaciones faltantes*. Aunque en situaciones como esta por lo común la forma de $p(\boldsymbol{\theta}|\boldsymbol{z})$ es bastante complicada, generalmente $p(\boldsymbol{\theta}|\boldsymbol{x})$ tiene una forma mucho más sencilla (esto ocurre, por ejemplo, en el caso de los diseños desbalanceados).

Notemos que

$$\begin{aligned} p(\boldsymbol{\theta}|\tilde{\boldsymbol{z}}, \boldsymbol{z}) &= p(\boldsymbol{\theta}|\boldsymbol{x}) \\ &\text{y} \\ p(\tilde{\boldsymbol{z}}|\boldsymbol{\theta}, \boldsymbol{z}) &= p(\tilde{\boldsymbol{z}}|\boldsymbol{\theta}). \end{aligned} \tag{2.10}$$

Por lo tanto podemos considerar a $\tilde{\boldsymbol{z}}$ como un "parámetro desconocido más" e incluirlo en un muestreo de Gibbs con densidades condicionales completas dadas precisamente por (2,10). De esta forma es posible generar observaciones de $p(\boldsymbol{\theta}|\boldsymbol{z})$, es decir, de la distribución que nos interesa. ◁

Convergencia

Supongamos que se desea generar una muestra de tamaño N de la distribución $p(\boldsymbol{\theta}|\boldsymbol{x})$. Si para cada uno de N valores iniciales $\boldsymbol{\theta}_1^{(0)}, ...\boldsymbol{\theta}_N^{(0)}$ corremos alguno de los algoritmos discutidos en esta sección, entonces, de acuerdo con la Proposición 2.1(i), después de un cierto número de iteraciones T suficientemente grande los valores $\boldsymbol{\theta}_1^{(T)}, ...\boldsymbol{\theta}_N^{(T)}$ pueden considerarse como una muestra de tamaño N de la distribución final de $\boldsymbol{\theta}$. Alternativamente podemos generar una sola cadena y tomar los valores $\boldsymbol{\theta}^{(T+k)}, \boldsymbol{\theta}^{(T+2k)}, ...\boldsymbol{\theta}^{(T+Nk)}$ como una muestra de $p(\boldsymbol{\theta}|\boldsymbol{x})$, donde k se elige de manera que la correlación entre las observaciones sea pequeña.

En general no es fácil determinar en qué momento la(s) cadena(s) ha(n) convergido. Un método empírico comúnmente utilizado, basado en la Proposición 2.1(ii), consiste en graficar los promedios ergódicos de algunas funciones de $\boldsymbol{\theta}$ contra el número de iteraciones y elegir el valor T a partir del cual las gráficas se estabilizan. En este caso es frecuente omitir los primeros valores de la(s) cadena(s) al calcular los promedios ergódicos. La idea de este *periodo de calentamiento* es permitir que la(s) cadena(s) salga(n) de una primera fase de inestabilidad. En el caso particular del muestreo de Gibbs, la velocidad de convergencia depende fuertemente de la correlación entre los componentes del vector $\boldsymbol{\theta}$ bajo la distribución final $p(\boldsymbol{\theta}|\boldsymbol{x})$: entre más alta sea la correlación más lenta será la convergencia.

2.2. Software

2.2.1. S-Plus y R

S-Plus

S-Plus es el programa estadístico comercial que se utiliza para el análisis clásico de los datos. Este programa utiliza el lenguaje S, un lenguaje creado para la exploración y visualización de los datos, modelación estadística y programación con datos. Trabaja en un ambiente de programación orientada a objetos.

Este programa está disponible en la página

$$http://www.insightful.com/products/default.html$$

R

R es un lenguaje y ambiente para cómputos estadísticos y gráficos. Es parte del proyecto GNU y utiliza el lenguaje y ambiente S, el cual fue desarrollado en los Laboratorios Bell (formalmente AT&T, ahora Lucent Technologies) por John Chambers y colegas. R puede ser considerado como una implementación diferente de S. Existen grandes diferencias, pero la mayor parte del código escrito para S corre inalterado en R.

R provee una amplia variedad de estadísticas (modelos lineales y no lineales, pruebas de estadística clásica, análisis de series de tiempo, etc.) y técnicas gráficas. R está disponible como software libre bajo los términos de Free Software Foundation.

El ambiente R:

R, como S, está diseñado alrededor de un lenguaje computacional y permite a los usuarios agregar funcionalidad definiendo nuevas funciones. La mayor parte del sistema está escrito en el dialecto R de S, el cual es fácil para los usuarios para seguir el algoritmo escogido. Para temas computacionales intensivos, los códigos en C, C++ y Fortran, pueden vincularse y correrse. Los usuarios avanzados pueden escribir en código C para manipular objetos de R directamente.

Este programa está disponible en la página de internet:

$$http://www.r-project.org$$

2.2.2. WinBUGS

El programa BUGS (Bayesian inference Using Gibbs Sampling) fue creado en 1989 como parte de un proyecto de investigación en la Unidad Bioestadística MRC del Instituto de Salud Pública Robinson Way en Cambridge; actualmente es desarrollado conjuntamente con el Departamento de Epidemiología y Salud Pública de la Escuela de Medicina del Imperial College en el Hospital St. Mary's de Londres.

BUGS es un programa que realiza inferencia bayesiana en problemas estadísticos complejos para los cuales no existe una solución analítica exacta y cualquier técnica de aproximación tiene dificultades. Las aplicaciones más comunes incluyen modelos lineales generalizados jerárquicos con efectos aleatorios, espaciales, temporales o transversales; problemas con datos faltantes; estimación restringida; y cualquier análisis en el que la información a priori deba ser incluida.

BUGS utiliza herramientas como métodos Monte Carlo vía cadenas de Markov (MCMC), generalmente utiliza el muestreo Gibbs univariado aunque puede utilizar rutinas sencillas del algoritmo Metropolis dentro del muestreo Gibbs cuando sea necesario. WINBUGS (la versión interactiva que trabaja en Windows) contiene un muestreo Metropolis más sofisticado y diagnósticos de convergencia. Ofrece una interfaz con el usuario basada en cuadros de diálogo y comandos a través de los cuales se analiza el modelo, por lo que el ambiente de WINBUGS se vuelve más amigable. Además también es posible realizar una interfaz con R o S-Plus, que manejan una sintaxis similar a la de WINBUGS.

BUGS proporciona un lenguaje para especificaciones de modelos estadísticos basados en la estrucutra gráfica, aunque existen algunas restricciones en la clase de modelos que pueden ser frecuentemente analizados. Un compilador procesa el modelo y los datos y finalmente elabora las distribuciones que posteriormente se utilizarán para el muestreo Gibbs. Finalmente, los algoritmos de muestreo apropiados son implementados para simular valores de las cantidades desconocidas en el modelo.

BOA (Bayesian Output Analysis Program) es un conjunto de funciones de S-Plus o R que contiene un rango amplio de diagnósticos y gráficas para representar los resultados de un análisis con MCMC y verificar la convergencia de la simulación. BOA realiza los diagnósticos de convergencia y los resultados estadísticos y gráficos para las muestras producidas por el muestreo Gibbs. Está basado en CODA (Convergence Diagnostic and Output Analysis software), sin embargo es más rápido, eficiente y flexible.

Es de hacer notar que WINBUGS simula un nodo a la vez; esto puede hacer la convergencia muy lenta y el programa ineficiente para modelos con parámetros altamente relacionados, tal como algunas estructuras de series de tiempo.

El programa está disponible en:

$$http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml$$

3. Regresiones aparentemente no relacionadas

Percy (1992) analiza modelos de regresión lineal multivariada de la forma

$$\tilde{\mathbf{y}}_i = \tilde{\mathbf{X}}_i\beta + \epsilon_{\mathbf{i}} \qquad \epsilon_{\mathbf{i}} \sim \text{ind}N_M[\mathbf{0}, \mathbf{\Sigma}] \tag{3.1}$$

donde $\tilde{\mathbf{y}}_i$ es un vector de respuestas $M \times 1$ y $\tilde{\boldsymbol{X}}_i$ es una matriz de variables explicativas $M \times K$ para el individuo $i = 1, \ldots, N$, β es un vector de coeficientes de regresión $K \times 1$, ϵ_i es un vector de residuales para el individuo i, "ind" implica independencia condicional de las variables y Σ es su matriz de covarianza. El vector $\tilde{\boldsymbol{y}}_i$ corresponde al i-ésimo renglón de la matriz $\boldsymbol{Y}$ descrita en la sección 2.1.3. Los coeficientes de regresión son usualmente distintos para cada variable de respuesta, por lo que $\tilde{\boldsymbol{X}}_i$ es de la forma:

$$\tilde{\boldsymbol{X}}_{\boldsymbol{i}} = \begin{pmatrix} \tilde{\boldsymbol{x}}_{\boldsymbol{i1}}^{\boldsymbol{T}} & \mathbf{0}^{\boldsymbol{T}} & \ldots & \mathbf{0}^{\boldsymbol{T}} \\ \mathbf{0}^{\boldsymbol{T}} & \tilde{\boldsymbol{x}}_{\boldsymbol{i2}}^{\boldsymbol{T}} & \ldots & \mathbf{0}^{\boldsymbol{T}} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}^{\boldsymbol{T}} & \mathbf{0}^{\boldsymbol{T}} & \ldots & \tilde{\boldsymbol{x}}_{\boldsymbol{iM}}^{\boldsymbol{T}} \end{pmatrix} \tag{3.2}$$

donde $\tilde{\boldsymbol{x}}_{ij}$ es un vector de variables explicativas $K_j \times 1$ para la respuesta j del individuo i y $K = \sum_{j=1}^{M} K_j$. Este vector corresponde al i-ésimo renglón de la matriz $\boldsymbol{X}_j$ descrita en la sección 2.1.3.

Zellner (1971) llamó a este modelo *Regresiones aparentemente no relacionadas* (SUR) ya que (3.1) puede ser escrito como M regresiones lineales múltiples separadas, las cuales contienen diferentes parámetros que están aparentemente no relacionados. Sin embargo, dichos parámetros están conectados porque las respuestas en las diferentes regresiones están correlacionadas entre sí. Press (1972) se refiere al modelo (3.1) como regresiones multivariadas generalizadas y Box & Tiao (1973) las llaman modelo lineal general multivariado.

El modelo (3.1) se reduce al modelo de regresión multivariada tradicional si $\tilde{\boldsymbol{x}}_{\boldsymbol{i1}} = \ldots = \tilde{\boldsymbol{x}}_{\boldsymbol{iM}}$ para $i = 1, \ldots, N$, lo que es conveniente cuando las variables explicativas son comunes a todas las respuestas para cada individuo. La densidad predictiva para este modelo puede ser evaluada explícitamente si se utiliza la distribución inicial no informativa de Jeffreys (Zellner y Chetty (1965) mostraron que es una t- Student multivariada para el vector futuro de respuestas). Se mostará la dificultad de evaluar la densidad predictiva bayesiana para el modelo SUR y se analizará una solución a este problema: el muestreo Gibbs. Finalmente, se discutirá una extensión al caso de datos faltantes en el vector de respuesta.

3.1. Inferencia bayesiana para el modelo de regresiones aparentemente no relacionadas

Zellner (1971), Press (1972) y Box & Tiao (1973) estudiaron la distribución final de los parámetros en el modelo SUR pero no consideraron el problema de predicción.

3.1.1. Distribución inicial y final

Como ya se había mencionado, por conveniencia, trabajaremos con la matriz de precisión $\Phi = \Sigma^{-1}$ más que con la matriz de covarianza Σ. En ausencia de información inicial la distribución inicial no informativa para β y Φ que utilizaremos será la distribución inicial de Jeffreys

$$f(\boldsymbol{\beta}, \Phi) \propto |\Phi|^{-(M+1)/2} \tag{3.3}$$

Del modelo (3.1), podemos definir el modelo SUR como

$$\tilde{\boldsymbol{y}}_i | \tilde{\boldsymbol{X}}_i, \boldsymbol{\beta}, \Phi \sim \text{ind} N_M[\tilde{\boldsymbol{X}}_i \boldsymbol{\beta}, \Phi^{-1}], \tag{3.4}$$

y la distribución conjunta final es entonces

$$\begin{aligned} & f(\beta, \Phi | \tilde{\boldsymbol{y}}_1, \ldots \tilde{\boldsymbol{y}}_N, \tilde{\boldsymbol{X}}_1, \ldots, \tilde{\boldsymbol{X}}_N) \\ & \quad \propto \quad |\Phi|^{(N-M-1)/2} \times \exp\Big\{ -\frac{1}{2} \sum_{i=1}^{N} (\tilde{\boldsymbol{y}}_i - \tilde{\boldsymbol{X}}_i \boldsymbol{\beta})^T \Phi (\tilde{\boldsymbol{y}}_i - \tilde{\boldsymbol{X}}_i \boldsymbol{\beta}) \Big\}. \end{aligned}$$

3.1.2. Densidad predictiva

Suponga ahora el modelo (3.4) para $i = 1, \ldots, N+1$ y que hemos observado $\tilde{\boldsymbol{X}}_1, \ldots \tilde{\boldsymbol{X}}_{N+1}$ y $\tilde{\boldsymbol{y}}_1, \ldots \tilde{\boldsymbol{y}}_N$. Queremos predecir $\tilde{\boldsymbol{y}}_{N+1}$ dados los datos observados. Sea $\mathbb{T} = \{\tilde{\boldsymbol{y}}_1, \ldots, \tilde{\boldsymbol{y}}_N, \tilde{\boldsymbol{X}}_1, \ldots, \tilde{\boldsymbol{X}}_N\}$ el conjunto de datos observados, por lo que la densidad predictiva será entonces:

$$\begin{aligned} & f(\tilde{\boldsymbol{y}}_{\boldsymbol{N+1}} | \tilde{\boldsymbol{X}}_{\boldsymbol{N+1}}, \mathbb{T}) \qquad (3.5) \\ & \quad \propto \quad \int_{\beta} \int_{\Phi} |\Phi|^{(N-M)/2} \exp\Big\{ -\frac{1}{2} \sum_{i=1}^{N+1} (\tilde{\boldsymbol{y}}_i - \tilde{\boldsymbol{X}}_i \boldsymbol{\beta})^T \Phi (\tilde{\boldsymbol{y}}_i - \tilde{\boldsymbol{X}}_i \boldsymbol{\beta}) \Big\} d\Phi d\boldsymbol{\beta}. \end{aligned}$$

Integrando con respecto a Φ tenemos

$$f(\tilde{\boldsymbol{y}}_{N+1} | \tilde{\boldsymbol{X}}_{N+1}, \mathbb{T}) \propto \int_{\beta} \Big| \sum_{i=1}^{N+1} (\tilde{\boldsymbol{y}}_i - \tilde{\boldsymbol{X}}_i \boldsymbol{\beta})(\tilde{\boldsymbol{y}}_i - \tilde{\boldsymbol{X}}_i \boldsymbol{\beta})^T \Big|^{-(N+1)/2} d\boldsymbol{\beta}$$

siempre que $N > M - 2$. Drèze (1977) mostró que esta integral es muy difícil de resolver analíticamente. Si cambiamos el orden de integración de la ecuación (3.5), esto da

$$f(\tilde{\boldsymbol{y}}_{N+1}|\tilde{\boldsymbol{X}}_{N+1},\mathbb{T})$$
$$\propto \int_{\Phi} |\Phi|^{(N-M)/2} \Big|\sum_{i=1}^{N+1}(\tilde{\boldsymbol{X}}_i^T \Phi \tilde{\boldsymbol{X}}_i)\Big|^{1/2} \times \exp\left(-\frac{1}{2}\frac{\Big|\sum_{i=1}^{N+1}(A_i^T\Phi A_i)\Big|}{\Big|\sum_{i=1}^{N+1}(\tilde{\boldsymbol{X}}_i^T\Phi\tilde{\boldsymbol{X}}_i)\Big|}\right) d\Phi$$

donde $A_i = (\tilde{\boldsymbol{y}}_i, \tilde{\boldsymbol{X}}_i)$ para $i = 1, \ldots, N+1$. El integrando es una función complicada de Φ en la que no se simplifica lo suficiente para facilitar la integración, a menos que $\tilde{\boldsymbol{x}}_{i1} = \ldots = \tilde{\boldsymbol{x}}_{iM}$ en la ecuación (3.2) para $i = 1, \ldots, N+1$.

3.2. Aproximaciones para la densidad predictiva

Gelfand et al. (1989) y Gelfand and Smith (1990) investigaron el algoritmo de muestreo Gibbs para estimar las funciones de densidad conjunta y marginales. Con base en la simulación Monte Carlo y usando distribuciones condicionales para actualizar la información de los estimadores de las interaciones desconocidas, se obtiene una ecuación explícita para la densidad estimada. El método es muy atractivo por la facilidad razonable de implementación.

Muestreo de Gibbs

Para el modelo SUR con la densidad inicial de Jeffreys, nuestras variables aleatorias no observadas son $\tilde{\boldsymbol{y}}_{N+1}, \Phi$ y β y deseamos calcular la función de densidad predictiva $f(\tilde{\boldsymbol{y}}_{N+1}|\tilde{\boldsymbol{X}}_{N+1},\mathbb{T})$. Entonces, las densidades condicionales completas que se requieren son:

$$f(\tilde{\boldsymbol{y}}_{N+1}|\tilde{\boldsymbol{X}}_{N+1},\boldsymbol{\beta},\Phi,\mathbb{T}) \equiv N_M[\tilde{\boldsymbol{X}}_{N+1}\boldsymbol{\beta},\Phi^{-1}], \tag{3.6}$$

$$f(\Phi|\tilde{\boldsymbol{X}}_{N+1},\tilde{\boldsymbol{Y}}_{N+1},\boldsymbol{\beta},\mathbb{T}) \equiv W_M\Big[N+1,\Big(\sum_{i=1}^{N+1}(\tilde{\boldsymbol{y}}_i - \tilde{\boldsymbol{X}}_i\boldsymbol{\beta})(\tilde{\boldsymbol{y}}_i - \tilde{\boldsymbol{X}}_i\boldsymbol{\beta})^T\Big)^{-1}\Big], \tag{3.7}$$

$$f(\boldsymbol{\beta}|\tilde{\boldsymbol{X}}_{N+1},\tilde{\boldsymbol{Y}}_{N+1},\Phi,\mathbb{T}) \tag{3.8}$$
$$\equiv N_q\Big[\Big(\sum_{i=1}^{N+1}(\tilde{\boldsymbol{X}}_i^T\Phi\tilde{\boldsymbol{X}}_i)\Big)^{-1}\Big(\sum_{i=1}^{N+1}(\tilde{\boldsymbol{X}}_i^T\Phi\tilde{\boldsymbol{y}}_i)\Big), \Big(\sum_{i=1}^{N+1}(\tilde{\boldsymbol{X}}_i^T\Phi\tilde{\boldsymbol{X}}_i)\Big)^{-1}\Big].$$

Para escalares fijos t y m, el algoritmo de Muestreo Gibbs procede de la manera siguiente:

(a) Selecciona valores iniciales arbitrarios, $\tilde{\boldsymbol{y}}_{N+1}^{(0)}$, $\Phi^{(0)}$ y $\boldsymbol{\beta}^{(0)}$. Sea $r = 0$.

(b) Simula $\tilde{\boldsymbol{y}}_{N+1}^{(r+1)}|\tilde{\boldsymbol{X}}_{N+1},\boldsymbol{\beta}^{(r)},\Phi^{(r)}$ de la densidad (3.6); simula $\Phi^{(r+1)}|\tilde{\boldsymbol{X}}_{N+1}$, $\tilde{\boldsymbol{y}}_{N+1}^{(r+1)}$, $\boldsymbol{\beta}^{(r)}$, $\mathbb{T}$ de la densidad (3.7); simula $\boldsymbol{\beta}^{(r+1)}|\tilde{\boldsymbol{X}}_{N+1},\tilde{\boldsymbol{y}}_{N+1}^{(r+1)}$, $\Phi^{(r+1)}$, $\mathbb{T}$ de la densidad (3.8).

(c) Sea $r = r + 1$ y repite el paso (b) hasta $r + 1 = t$. $\boldsymbol{\beta}^{(t)}$ y $\Phi^{(t)}$ son parámetros estimados de $f(\tilde{\boldsymbol{y}}_{N+1}|\tilde{\boldsymbol{X}}_{N+1}, \boldsymbol{\beta}, \Phi)$.

(d) Repite los pasos (a) a (c) hasta que se hayan ejecutado m ciclos y se hayan obtenido m pares de parámetros estimados : $\boldsymbol{\beta}^{(t)}_{(1)}, \ldots, \boldsymbol{\beta}^{(t)}_{(m)}$ y $\Phi^{(t)}_{(1)}, \ldots, \Phi^{(t)}_{(m)}$.

La aproximación de la densidad predictiva está dada por

$$f(\tilde{\boldsymbol{y}}_{N+1}|\tilde{\boldsymbol{X}}_{N+1}, \mathbb{T}) \simeq \frac{1}{m}\sum_{j=1}^{m} f(\tilde{\boldsymbol{y}}_{N+1}|\tilde{\boldsymbol{X}}_{N+1}, \beta^{(t)}_{(j)}), \Phi^{(t)}_{(j)}, \mathbb{T}) \tag{3.9}$$

la cual es evaluada utilizando la expresión (3.6).

Geman and Geman (1984) mostraron que el algoritmo Gibbs converge a la densidad predictiva verdadera cuando $t \to \infty$.

3.3. Extensión para analizar datos faltantes

Si algunos de los componentes de $\tilde{\boldsymbol{y}}_i$ y $\tilde{\boldsymbol{X}}_i$ son no observados, debemos verlos como variables aleatorias. Por lo tanto, se requieren supuestos distribucionales adicionales y las integrales que resultan son generalmente muy laboriosas. Little y Rubin (1987) discuten un algoritmo apropiado que puede ser utilizado en estas circunstancias.

Sin embargo, si los datos faltantes ocurren sólo en el vector de respuesta, la integral múltiple que conduce a la función de densidad predictiva no es tan complicada y podemos aplicar las aproximaciones para esta expresión, de la siguiente manera.

Habíamos especificado anteriormente sólo tres distribuciones condicionales completas; ahora especificaremos una cuarta:

$$\begin{aligned} & f(\tilde{\boldsymbol{y}}^{(1)}_1, \ldots, \tilde{\boldsymbol{y}}^{(1)}_N | \tilde{\boldsymbol{X}}_1, \ldots, \tilde{\boldsymbol{X}}_{N+1}, \tilde{\boldsymbol{y}}^{(0)}_1, \ldots, \tilde{\boldsymbol{y}}^{(0)}_N, \tilde{\boldsymbol{y}}_{N+1}, \beta, \Phi) \\ & \quad \propto \prod_{i=1}^{N} f(\tilde{\boldsymbol{y}}^{(1)}_i | \tilde{\boldsymbol{y}}^{(0)}_i, \tilde{\boldsymbol{X}}_i, \beta, \Phi) \end{aligned}$$

donde $\tilde{\boldsymbol{y}}^{(0)}_i$ y $\tilde{\boldsymbol{y}}^{(1)}_i$ contienen los componentes observados y no observados de $\tilde{\boldsymbol{y}}_i$ respectivamente. Las funciones de densidad de probabilidad condicional en el producto son normales multivariadas y pueden obtenerse mediante el uso de resultados estándar de las distribuciones condicionales de subconjuntos de las variables normales multivariadas. Ahora es relativamente simple incluir esto en el algoritmo de muestreo de Gibbs dado anteriormente.

3.4. Ejemplo

Considere un vector de respuesta univariado pero con coeficientes de regresión diferentes para cada ecuación y suponga que existe información inicial

que sugiere correlación entre los coeficientes de las diferentes muestras, por ejemplo, si el modelo se refiere a un conjunto de empresas que enfrentan las mismas condiciones económicas. Suponga que tenemos datos temporales para N años y para M empresas relativos a los niveles brutos de inversión y_{ij}, $(i = 1, ..., N; j = 1, ..., M)$. Consideremos las variaciones anuales en inversión con relación al capital de la empresa C_{ij} al principio de cada año, y al inicio del año del mercado de valores V_{ij} de las firmas. La relación para cada una de las empresas será la forma:

$$\begin{aligned} y_{i1} &= \alpha_1 + \beta_1 V_{i1} + \gamma_1 C_{i1} + \epsilon_{i1} \\ y_{i2} &= \alpha_2 + \beta_2 V_{i2} + \gamma_2 C_{i2} + \epsilon_{i2} \\ &\cdots \\ y_{iM} &= \alpha_M + \beta_M V_{iM} + \gamma_M C_{iM} + \epsilon_{iM} \end{aligned}$$

Los errores en cada conjunto de ecuaciones pueden tomarse como no correlacionados entre sí con $\epsilon_{i1} \sim N(0, \theta_1), \epsilon_{i2} \sim N(0, \theta_2), \ldots, \epsilon_{iM} \sim N(0, \theta_M)$. De esta forma, para $M = 2$ y $N = 10$, tendremos,

$$\boldsymbol{y}_j = \boldsymbol{X}_j \boldsymbol{\delta}_j + \boldsymbol{\epsilon}_j,$$

donde $\boldsymbol{y}_j$ y $\boldsymbol{\epsilon}_j$ son vectores de dimensión 10×1 para $j = 1, 2$. La matriz $\boldsymbol{X}_j$ tiene renglones $(1, V_{ij}, C_{ij})$ y $\boldsymbol{\delta}_j = (\alpha_j, \beta_j, \gamma_j)$ para $j = 1, 2$. De manera equivalente,

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\delta} + \boldsymbol{\epsilon}$$

donde y es 20×1, y X es una matriz diagonal 20×6,

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 & 0 \\ 0 & \boldsymbol{X}_2 \end{bmatrix}$$

con submatrices $\boldsymbol{X}_i$ (10×3) que contienen covariables para la empresa i, y $\boldsymbol{\delta}$ es un vector 6×1 con $\boldsymbol{\delta}' = (\alpha_1, \beta_1, \gamma_1, \alpha_2, \beta_2, \gamma_2)$. Con el supuesto de que no existe correlación entre los dos conjuntos de ecuaciones

$$\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \end{bmatrix} \sim N_2 \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \theta_1 \boldsymbol{I}_{10} & \boldsymbol{0}_{10} \\ \boldsymbol{0}_{10} & \theta_2 \boldsymbol{I}_{10} \end{bmatrix} \right) \tag{3.10}$$

donde $\boldsymbol{I}_{10}$ y $\boldsymbol{0}_{10}$ son matrices identidad y nula 10×10, respectivamente. La estimación por separado de cada ecuación ignora las posibles correlaciones de los errores a través de las ecuaciones. Las correlaciones de los errores pueden especificarse a través de un modelo de la forma

$$\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \end{bmatrix} \sim N_2 \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \theta_{11} \boldsymbol{I}_{10} & \theta_{12} \boldsymbol{I}_{10} \\ \theta_{12} \boldsymbol{I}_{10} & \theta_{22} \boldsymbol{I}_{10} \end{bmatrix} \right)$$

Theil (1971) tabuló los valores de las acciones, el capital y los niveles de inversión de dos de las empresas más importantes de Estados Unidos para el periodo 1934-53, llamadas General Electric y Westinghouse. Los datos están en millones de dólares a precios corrientes. El método de máxima verosimilitud se aplicó por separado a cada conjunto de datos como en el modelo (3.10), produciendo los siguientes estimadores (errores estándar en paréntesis):

$$\begin{aligned} y_{i1} &= -9{,}96 + 0{,}027V_{i1} + 0{,}152C_{i1} \\ &\quad (31{,}4) \qquad (0{,}015) \qquad (0{,}026) \\ y_{i2} &= -0{,}51 + 0{,}053V_{i2} + 0{,}092C_{i2} \\ &\quad (8{,}0) \qquad (0{,}016) \qquad (0{,}050) \end{aligned}$$

Al tomar en cuenta la correlación entre dos conjuntos de ecuaciones mejora la precisión de los coeficientes estimados de V pero disminuye la significancia del coeficiente para C_2. La matriz de dispersión conjunta de ϵ_1 y ϵ_2 está dada por

$$\Sigma \otimes I_{10} = \begin{bmatrix} 824 & 230 \\ 230 & 114 \end{bmatrix} \otimes I_{10}$$

Los resultados para los parámetros restantes, basados en una corrida de 10 000 iteraciones con 500 iteraciones de "calentamiento"se muestran en la siguiente tabla:

Parámetro	**Media**	**Desviación estándar**	**2.5 %**	**Mediana**	**97.5 %**
α_1	-16.4	22.5	-60.2	-16.9	28.0
β_1	0.034	0.012	0.010	0.034	0.060
γ_1	0.133	0.026	0.078	0.134	0.183
α_2	0.350	7.158	-13.660	0.409	13.730
β_2	0.056	0.015	0.028	0.056	0.086
γ_2	0.059	0.059	-0.058	0.059	0.176

El código utilizado en WINBUGS para realizar este ejercicio fue:

```
PROGRAM

model {
for (t in 1:T) {
y[t,1:2] ~ dmnorm(mu[t,],Omega[ , ])

# means in separate time series
     mu[t,1] <- beta[1,1]+beta[1,2]*v[t,1]+beta[1,3]*k[t,1]
     mu[t,2] <- beta[2,1]+beta[2,2]*v[t,2]+beta[2,3]*k[t,2]
```

```
}

# priors on regression coefficients
for (i in 1:M) {
for (j in 1:3) {beta[i,j] ~ dnorm(0,0.001)
}
}

Omega[1 : M , 1 : M]  ~ dwish(R[ , ], 2)

# cross-correlation matrix of dimension M=2
#for (i in 1 : M) {
# for (j in 1 : M) {
# Sigma[i, j] <- inverse(Omega[ , ], i, j)
Sigma[1:M,1:M] <- inverse(Omega[1:M, 1:M])
# }
#}

rho.sq <- Sigma[1,2]*Sigma[1,2]/(Sigma[1,1]*Sigma[2,2])
}

DATA 1
list(T=20,M=2,R = structure(.Data = c(1, 0, 0, 1),.Dim = c(2, 2)))

DATA 2
y[,1] v[,1] k[,1] y[,2] v[,2] k[,2]
33.1 1170.6 97.8 12.93 191.5 1.8
45 2015.8 104.4 25.9 516 0.8
77.2 2803.3 118 35.05 729 7.4
44.6 2039.7 156.2 22.89 560.4 18.1
48.1 2256.2 172.6 18.84 519.9 23.5
74.4 2132.2 186.6 28.57 628.5 26.5
113 1834.1 220.9 48.51 537.1 36.2
91.9 1588 287.8 43.34 561.2 60.8
61.3 1749.4 319.9 37.02 617.2 84.4
56.8 1687.2 321.3 37.81 626.7 91.2
93.6 2007.7 319.6 39.27 737.2 92.4
159.9 2208.3 346 53.46 760.5 86
147.2 1656.7 456.4 55.56 581.4 111.1
146.3 1604.4 543.4 49.56 662.3 130.6
98.3 1431.8 618.3 32.04 635.2 136.7
```

```
93.5 1610.5 647.4 32.24 635.2 136.7
135.2 1819.4 671.3 54.38 723.8 129.7
157.3 2079.7 726.1 71.78 864.1 145.5
179.5 2371.6 800.3 90.08 1193.5 174.8
189.6 2759.9 888.9 68.6 1188.9 213.5
END

Inits
list(beta=structure(.Data=c(0,0,0,0,0,0),.Dim=c(2,3)))
```

En el siguiente capítulo se utilizará un conjunto de datos reales relativos a diversos sectores de la actividad económica en México medidos a través del PIB de 1990 a 2006.

4. Aplicación

El modelo SUR es un método estadístico que tiene varias aplicaciones en series de tiempo, principalmente en econometría ya que, como hemos visto, realiza regresiones longitudinales de datos observados de cualquier evento; es decir, permite estudiar eventos que ocurren a través del tiempo.

Generalmente un evento económico se define como un cambio cualitativo que ocurre en un punto específico en el tiempo y dicho cambio debe consistir en una diferencia significativa entre lo que precede y lo que sigue. Como los eventos están definidos en términos de cambios a través del tiempo, la mejor forma de estudiar estos eventos y sus causas es recolectando un historial de los eventos bajo estudio. En su forma más simple, un historial de este tipo es un registro longitudinal de los eventos sucedidos en una muestra.

Para realizar este análisis deben tomarse en cuenta aspectos tales como la definición del evento, el tiempo de observación del evento y principalmente los supuestos estadísticos. Para los periodos de recopilación de datos es recomendable tener una observación en cada fase de una sucesión de eventos para asegurar un estudio adecuado.

Un ejemplo sobre este modelo consiste en el cálculo del PIB en México. En este caso, tenemos que el PIB Total está determinado por el valor del PIB por División en cada periodo (en este caso datos trimestrales). Es razonable suponer que los errores aleatorios asociados con las ecuaciones de cada sector pueden estar correlacionados temporalmente, dada la presencia de la economía nacional común influyendo en cada una de las divisiones. En particular, si el término de error en la primera ecuación refleja (al menos en una parte) la omision de algunas variables no observables; entonces estas mismas variables, u otras que estén altamente correlacionadas con ellas, pueden ser determinantes importantes de la variabilidad del error en la otra ecuación. Por lo tanto, las ecuaciones de cada División son regresiones aparentemente no relacionadas.

4.1. Producto Interno Bruto (PIB) de México

El Producto Interno Bruto es el valor total de la producción de bienes y servicios finales generados dentro del territorio nacional durante un cierto periodo de tiempo.

El PIB está conformado por nueve divisiones: en primer lugar se tiene la División IX *Servicios Comunales, Sociales y Personales* que representa un 22.6 % del PIB, seguida por las divisiones VI *Comercio, Restaurantes y Hoteles* y la III *Industria Manufacturera* con una participación de 19.1 % y 17.5 %, respectivamente. Las divisiones VIII *Servicios Financieros, Seguros, Actividades Inmobiliarias y de Alquiler*, VII *Transporte, Almacenaje y Comunicaciones*, IV *Construcción* y I *Agropecuario, Silvicultura y Pesca* en su conjunto apenas con-

tribuyen con el 30 % del PIB y por último, se encuentran las divisiones cuya participación es menor al 1.5 %, la División II: *Minería* y V *Electricidad, Gas y Agua*. Es importante mencionar que la participación de cada sector de actividad en el PIB Total depende principalmente de la productividad, es decir, de la relación entre la producción de bienes, en el caso de una empresa manufacturera, o ventas en el de los servicios, y las cantidades de insumos utilizados. De esta manera, el concepto de productividad es igualmente aplicable a una empresa industrial o de servicios, a un comercio, a una industria o al agregado de la economía. En pocas palabras, la productividad nos indica cuánto producto generan los insumos utilizados en una actividad económica.

4.2. Cálculo del PIB

Para mostrar un ejemplo sencillo sobre el modelo SUR aplicado al cálculo del PIB, se eligieron sólo las siguientes divisiones (I a VI): Agropecuario, silvicultura y pesca; Minería; Manufacturas; Construcción; Electricidad, gas y agua; Comercio, restaurantes y hoteles; ya que son las divisiones con facilidad de acceso a sus bases de datos. En cada división se eligieron las variables con alta correlación con su respectiva variable de respuesta y con los datos completos para el periodo establecido (1990-2006). En cambio, se eliminaron las variables altamente correlacionadas con variables explicativas de otras divisiones. Al final, las variables utilizadas para la regresión del PIB por división son mostradas en las siguientes tablas:

I PIB Agropecuario, silvicultura y pesca

Variable	Nombre
I.1	PIB Agropecuario
I.3	Producción de ganadería
I.5	Producción pesquera nacional

II PIB Minería

Variable	Nombre
II.1	PIB Minería
II.2	Volumen de producción de metales preciosos
II.3	Volumen de producción de metales industriales no ferrosos
II.4	Volumen de producción de metales y minerales siderúrgicos
II.5	Indice de producción minerometalúrgica

III PIB Manufacturas

Variable	Nombre
III.1	PIB Manufacturas
III.4	Personal ocupado en maquila
III.5	Valor agregado de exportación (cobrado por el servicio de maquila)

IV PIB Construcción

Variable	Nombre
IV.1	PIB Construcción
IV.2	Valor de producción de la construcción
IV.4	Valor de producción de la construcción privada

V PIB Electricidad, gas y agua

Variable	Nombre
V.1	PIB Electricidad, gas y agua
V.2	Generación CFE
V.6	Volumen almacenado en las principales presas de México

VI PIB Comercio, restaurantes y hoteles

Variable	Nombre
VI.1	PIB Comercio, restaurantes y hoteles
VI.4	Gasto de visitantes internacionales a México

4.3. Modelo SUR utilizado para el PIB

El modelo se refiere a un conjunto de actividades productivas en México llamadas *divisiones* que enfrentan las mismas condiciones económicas. Los datos están disponibles para $N = 16$ años y para $M = 6$ divisiones, donde cada y_{ij} representa el nivel de producción en términos del PIB, $(i = 1, ..., N; j = 1, ..., M)$. Cada división considera diferentes variables para ajustar su regresión, por lo que la relación de las divisiones será de la forma,

$$\begin{aligned} y_{i1} &= \boldsymbol{x'}_{i1}\boldsymbol{\beta}_1 + \epsilon_{i1} \\ y_{i2} &= \boldsymbol{x'}_{i2}\boldsymbol{\beta}_2 + \epsilon_{i2} \\ &\cdots \\ y_{iM} &= \boldsymbol{x'}_{iM}\boldsymbol{\beta}_M + \epsilon_{iM} \end{aligned}$$

donde $\boldsymbol{x}_{ij}$ es un vector de $1 \times k_j$ y $\boldsymbol{\beta}_j$ es un vector de $k_j \times 1$, con k_j correspondientes al número de variables utilizadas para ajustar el modelo de la división j más un uno en la primera entrada. Los errores en cada conjunto de ecuaciones pueden tomarse como no correlacionados entre sí con $\epsilon_{i1} \sim N(0, \theta_1), \epsilon_{i2} \sim N(0, \theta_2), \ldots, \epsilon_{iM} \sim N(0, \theta_M)$. De esta forma, para $M = 6$ y $N = 16$, tendremos,

$$\boldsymbol{y}_j = \boldsymbol{X}_j \boldsymbol{\beta}_j + \boldsymbol{\epsilon}_j,$$

donde $\boldsymbol{y}_j$ y $\boldsymbol{\epsilon}_j$ son vectores de dimensión 16×1 para $j = 1, \ldots 6$. La matriz $\boldsymbol{X}_j$ tiene renglones $(\boldsymbol{x_{ij}})$ para $j = 1, ..., 6$. De manera equivalente,

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

donde $\boldsymbol{y}$ es 96×1, y $\boldsymbol{X}$ es una matriz diagonal $96 \times \sum_{j=1}^{N} k_j$,

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & \boldsymbol{X}_2 & 0 & 0 & 0 & 0 \\ 0 & 0 & \boldsymbol{X}_3 & 0 & 0 & 0 \\ 0 & 0 & 0 & \boldsymbol{X}_4 & 0 & 0 \\ 0 & 0 & 0 & 0 & \boldsymbol{X}_5 & 0 \\ 0 & 0 & 0 & 0 & 0 & \boldsymbol{X}_6 \end{bmatrix}$$

con submatrices $\boldsymbol{X}_j$ $(16 \times k_j)$ que contienen covariables para la división j, y $\boldsymbol{\beta}$ es un vector $\sum_{i=1}^{N} k_i \times 1$ con $\boldsymbol{\beta}' = (\beta_1, \beta_2, ..., \beta_M)$. Con el supuesto de que no existe correlación entre los ocho conjuntos de ecuaciones

$$\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_M \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} \theta_1 \boldsymbol{I}_{16} & \boldsymbol{0}_{16} & \ldots & \boldsymbol{0}_{16} \\ \boldsymbol{0}_{16} & \theta_2 \boldsymbol{I}_{16} & \ldots & \boldsymbol{0}_{16} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0}_{16} & \boldsymbol{0}_{16} & \ldots & \theta_M \boldsymbol{I}_{16} \end{bmatrix} \right) \tag{4.1}$$

donde $\boldsymbol{I}_{16}$ y $\boldsymbol{0}_{16}$ son matrices identidad y nula 16×16, respectivamente. La estimación por separado de cada ecuación ignora las posibles correlaciones de los errores a través de las ecuaciones.

Las correlaciones de los errores pueden especificarse a través de un modelo SUR de la forma

$$\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_M \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} \theta_{11} \boldsymbol{I}_{16} & \theta_{12} \boldsymbol{I}_{10} & \ldots & \theta_{1M} \boldsymbol{I}_{16} \\ \theta_{12} \boldsymbol{I}_{16} & \theta_{22} \boldsymbol{I}_{16} & \ldots & \theta_{2M} \boldsymbol{I}_{16} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{1M} \boldsymbol{I}_{16} & \theta_{22} \boldsymbol{I}_{16} & \ldots & \theta_{MM} \boldsymbol{I}_{16} \end{bmatrix} \right)$$

Los resultados obtenidos por máxima verosimilitud suponiendo independencia entre las regresiones son:

$$
\begin{aligned}
y_{i1} &= 34{,}75 + 1{,}684(I{,}3_i) + 7{,}34(I{,}5_i) \\
&\quad (4{,}198) \qquad (0{,}1187) \qquad (3{,}333) \\
y_{i2} &= 6{,}372 - 30{,}72(II{,}2_i) - 62{,}68(II{,}3_i) - 0{,}6759(II{,}4_i) + 0{,}2085(II{,}5_i) \\
&\quad (5{,}939) \qquad (35{,}67) \qquad (152{,}2) \qquad (4{,}222) \qquad (0{,}1033) \\
y_{i3} &= 156{,}2 + 65{,}36(III{,}4_i) + 4{,}612(III{,}5_i) \\
&\quad (6{,}604) \qquad (21{,}21) \qquad (1{,}194) \\
y_{i4} &= 34{,}74 + 2{,}379(IV{,}2_i) + 7{,}73(IV{,}4_i) \\
&\quad (6{,}962) \qquad (2{,}036) \qquad (7{,}009) \\
y_{i5} &= -0{,}846 + 0{,}1424(V{,}2_i) - 0{,}005615(V{,}6_i) \\
&\quad (1{,}063) \qquad (0{,}003895) \qquad (0{,}01142) \\
y_{i6} &= 123{,}7 + 20{,}46(VI{,}4_i) \\
&\quad (16{,}77) \qquad (2{,}048)
\end{aligned}
$$

Por su parte, los resultados obtenidos a partir del modelo SUR se muestran en las siguientes ecuaciones,

$$
\begin{aligned}
y_{i1} &= 35{,}07 + 1{,}683(I{,}3_i) + 7{,}142(I{,}5_i) \\
&\quad (4{,}758) \qquad (0{,}2181) \qquad (2{,}21) \\
y_{i2} &= 8{,}566 - 39{,}66(II{,}2_i) + 142{,}4(II{,}3_i) + 3{,}213(II{,}4_i) + 0{,}03334(II{,}5_i) \\
&\quad (2{,}8) \qquad (21{,}5) \qquad (66{,}66) \qquad (1{,}632) \qquad (0{,}04148) \\
y_{i3} &= 164{,}4 + 106{,}5(III{,}4_i) + 0{,}8429(III{,}5_i) \\
&\quad (6{,}901) \qquad (11{,}63) \qquad (0{,}7631) \\
y_{i4} &= 54{,}4 + 0{,}918(IV{,}2_i) + 0{,}0509(IV{,}4_i) \\
&\quad (3{,}016) \qquad (0{,}7737) \qquad (2{,}044) \\
y_{i5} &= -0{,}3402 + 0{,}1392(V{,}2_i) - 0{,}00486(V{,}6_i) \\
&\quad (2{,}164) \qquad (0{,}01171) \qquad (0{,}01504) \\
y_{i6} &= 224{,}3 + 7{,}791(VI{,}4_i) \\
&\quad (24{,}57) \qquad (2{,}946)
\end{aligned}
$$

Los valores de los parámetros obtenidos a partir del muestreo Gibbs se encuentran dentro de los intervalos de confianza correspondientes; sin embargo, bajo el modelo SUR disminuye la longitud del intervalo para la mayor parte de los casos.

Los datos obtenidos al realizar las regresiones se muestran en las siguientes tablas.

Al hacer las regresiones por separado para cada división, se obtuvo:

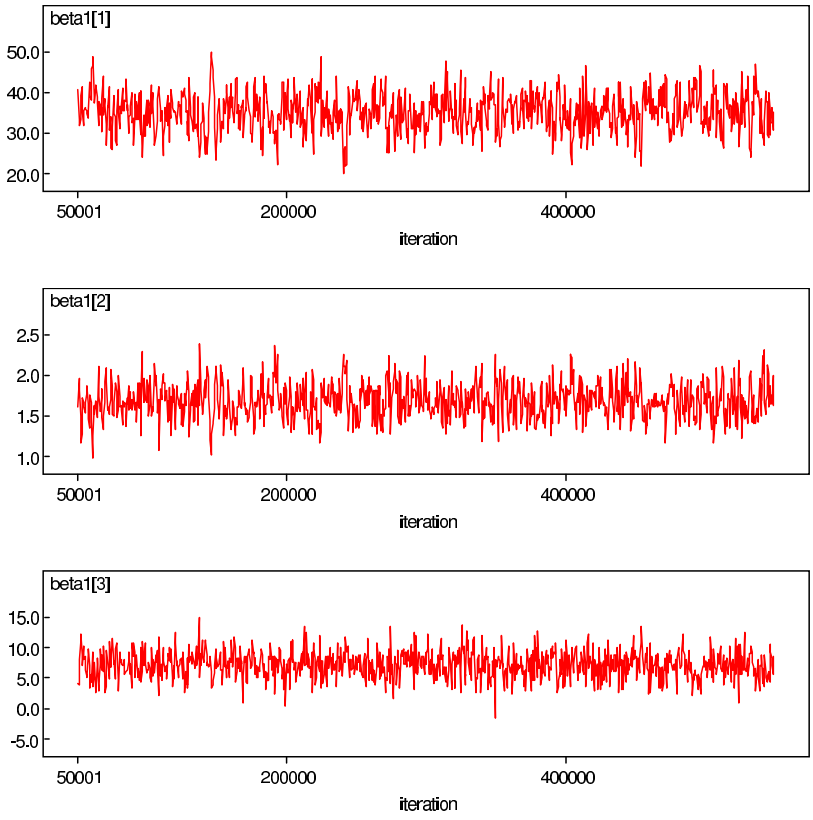
Parámetro	Media	Desv. est.	ECM	2.5 %	Mediana	97.5 %
beta1[1]	34.75	4.198	0.04096	26.36	34.78	42.97
beta1[2]	1.684	0.1187	0.001101	1.452	1.684	1.918
beta1[3]	7.34	3.333	0.03166	0.7709	7.288	14.03
beta2[1]	6.372	5.939	0.05199	-5.501	6.435	18.03
beta2[2]	-30.72	35.67	0.3414	-102.4	-30.35	39.24
beta2[3]	-62.68	152.2	1.492	-365.4	-63.13	245.4
beta2[4]	-0.6759	4.222	0.04484	-8.918	-0.665	7.758
beta2[5]	0.2085	0.1033	0.001029	0.004611	0.2092	0.4086
beta3[1]	156.2	6.604	0.07767	142.9	156.2	169.3
beta3[2]	65.36	21.21	0.2347	24.14	65.22	107.4
beta3[3]	4.612	1.194	0.0126	2.254	4.616	6.968
beta4[1]	34.74	6.962	0.07004	21.02	34.71	48.28
beta4[2]	2.379	2.036	0.02069	-1.648	2.392	6.443
beta4[3]	7.73	7.009	0.07002	-6.132	7.732	21.69
beta5[1]	-0.846	1.063	0.009339	-2.966	-0.8443	1.236
beta5[2]	0.1424	0.003895	3.43E-05	0.1347	0.1424	0.1502
beta5[3]	-0.005615	0.01142	9.18E-05	-0.02851	-0.005743	0.01725
beta6[1]	123.7	16.77	0.1746	89.44	123.7	156.9
beta6[2]	20.46	2.048	0.02072	16.37	20.47	24.57
sigma[1]	2.289	1.021	0.01272	1.066	2.061	4.855
sigma[2]	1.672	0.8011	0.01053	0.7256	1.486	3.747
sigma[3]	45.3	19.91	0.2117	21.03	40.86	96.11
sigma[4]	30.11	13.4	0.1366	13.95	27.05	64.58
sigma[5]	0.2425	0.1061	0.001282	0.1124	0.2179	0.5106
sigma[6]	291.4	124.4	1.294	137.7	265.7	614.6

Los resultados de los parámetros a partir del modelo SUR, basados en una corrida de 500,000 iteraciones se muestran en la siguiente tabla:

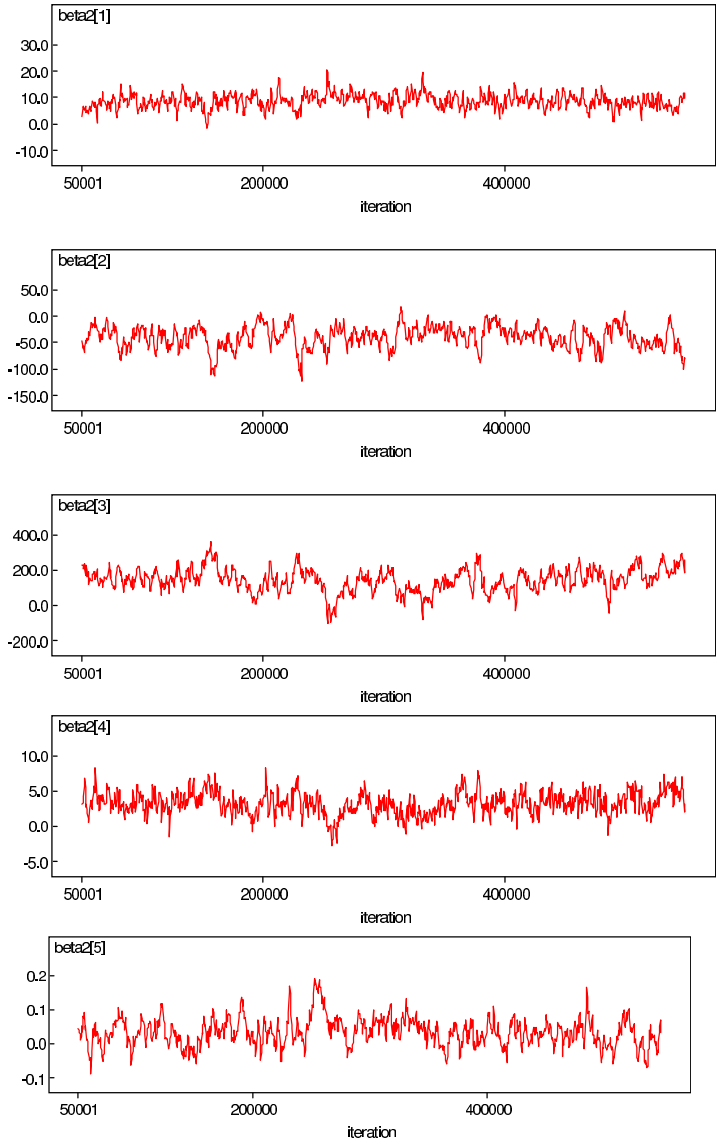
Parámetro	**Media**	**Desv. est.**	**ECM**	**2.5 %**	**Mediana**	**97.5 %**
beta1[1]	35.07	4.758	0.2366	25.36	35.2	43.96
beta1[2]	1.683	0.2181	0.009451	1.276	1.668	2.125
beta1[3]	7.142	2.21	0.09252	2.864	7.173	11.63
beta2[1]	8.566	2.8	0.2239	3.067	8.582	14.05
beta2[2]	-39.66	21.5	1.945	-85.2	-37.92	-2.114
beta2[3]	142.4	66.66	8.067	9.858	144.8	278.5
beta2[4]	3.213	1.632	0.1618	0.03802	3.179	6.324
beta2[5]	0.03334	0.04148	0.004134	-0.04628	0.03228	0.1313
beta3[1]	164.4	6.901	0.3	151.1	164.3	178.1
beta3[2]	106.5	11.63	0.6973	86.24	105.7	131.9
beta3[3]	0.8429	0.7631	0.04517	-0.779	0.8824	2.261
beta4[1]	54.4	3.016	0.1396	48.14	54.59	60.3
beta4[2]	0.918	0.7737	0.04298	-0.4854	0.8707	2.522
beta4[3]	0.0509	2.044	0.09531	-4.234	0.1084	3.74
beta5[1]	-0.3402	2.164	0.1217	-4.551	-0.4659	4.028
beta5[2]	0.1392	0.01171	6.88E-04	0.1154	0.1395	0.1628
beta5[3]	-0.00486	0.01504	5.13E-04	-0.03384	-0.00489	0.02627
beta6[1]	224.3	24.57	1.389	177.2	224.2	272.5
beta6[2]	7.791	2.946	0.1758	2.233	7.799	13.66

Parámetro	Media	Desv. est.	ECM	2.5 %	Mediana	97.5 %
Sigma[1,1]	2.633	1.495	0.05015	1.035	2.272	6.708
Sigma[1,2]	0.1098	1.229	0.04862	-2.084	0.09363	2.66
Sigma[1,3]	5.255	9.358	0.3406	-10.4	4.86	25.32
Sigma[1,4]	2.874	5.719	0.2376	-7.433	2.623	14.88
Sigma[1,5]	0.113	0.4595	0.0226	-0.7083	0.0745	1.1
Sigma[1,6]	-0.1584	29.03	1.216	-57.8	0.952	56.73
Sigma[2,1]	0.1098	1.229	0.04862	-2.084	0.09363	2.66
Sigma[2,2]	2.087	0.9302	0.03368	0.9141	1.879	4.408
Sigma[2,3]	13.79	8.031	0.3612	4.322	12.24	35.28
Sigma[2,4]	7.353	3.718	0.1436	2.338	6.721	16.99
Sigma[2,5]	0.04299	0.5497	0.02898	-1.04	0.04131	1.243
Sigma[2,6]	38.94	20.05	0.7271	12.98	35.25	94.48
Sigma[3,1]	5.255	9.358	0.3406	-10.4	4.86	25.32
Sigma[3,2]	13.79	8.031	0.3612	4.322	12.24	35.28
Sigma[3,3]	128.7	81.98	3.636	42.76	108.6	333.8
Sigma[3,4]	70.57	35.07	1.397	28.22	62.71	161.1
Sigma[3,5]	0.2186	4.214	0.2244	-8.19	0.2271	8.805
Sigma[3,6]	299.7	178.4	7.462	89.98	259	725.9
Sigma[4,1]	2.874	5.719	0.2376	-7.433	2.623	14.88
Sigma[4,2]	7.353	3.718	0.1436	2.338	6.721	16.99
Sigma[4,3]	70.57	35.07	1.397	28.22	62.71	161.1
Sigma[4,4]	46.75	19.28	0.7032	22.52	42.64	93.68
Sigma[4,5]	0.5349	2.526	0.1468	-4.368	0.4037	6.252
Sigma[4,6]	195.6	91.29	3.49	73.94	181.6	417.3
Sigma[5,1]	0.113	0.4595	0.0226	-0.7083	0.0745	1.1
Sigma[5,2]	0.04299	0.5497	0.02898	-1.04	0.04131	1.243
Sigma[5,3]	0.2186	4.214	0.2244	-8.19	0.2271	8.805
Sigma[5,4]	0.5349	2.526	0.1468	-4.368	0.4037	6.252
Sigma[5,5]	0.4461	0.3289	0.01433	0.1637	0.3509	1.293
Sigma[5,6]	4.174	13.4	0.7394	-20.82	3.227	36.64
Sigma[6,1]	-0.1584	29.03	1.216	-57.8	0.952	56.73
Sigma[6,2]	38.94	20.05	0.7271	12.98	35.25	94.48
Sigma[6,3]	299.7	178.4	7.462	89.98	259	725.9
Sigma[6,4]	195.6	91.29	3.49	73.94	181.6	417.3
Sigma[6,5]	4.174	13.4	0.7394	-20.82	3.227	36.64
Sigma[6,6]	1016	542.8	21.69	322.9	907.5	2418

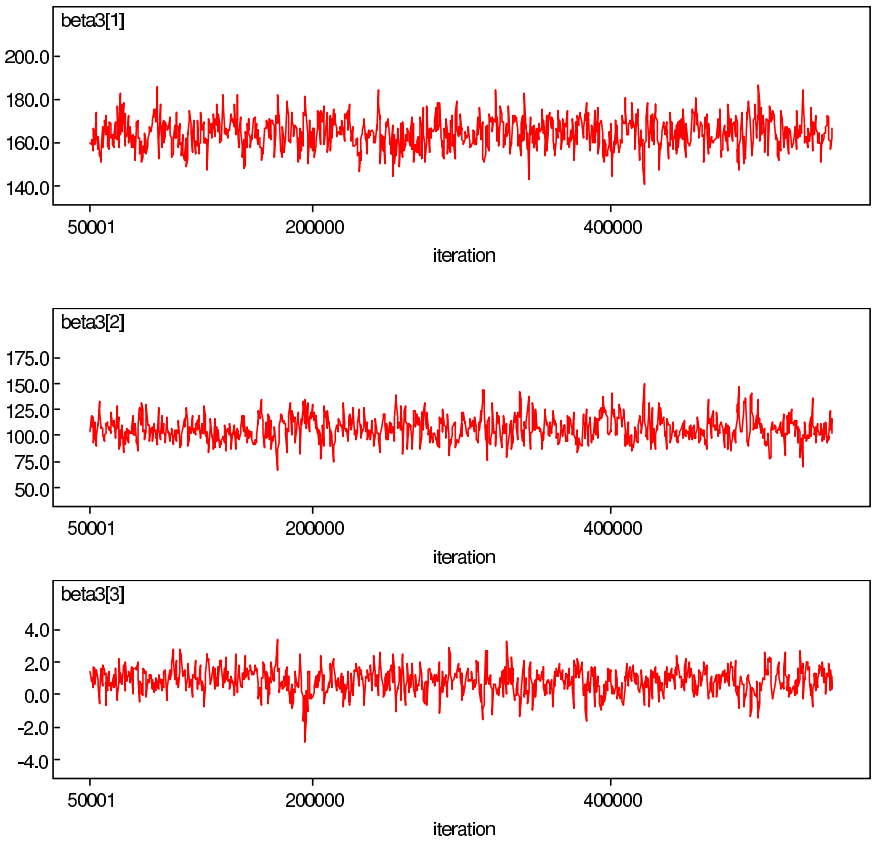
Gráficamente, la convergencia de las cadenas se muestran en las siguientes figuras. Los parámetros correspondientes a la división I:



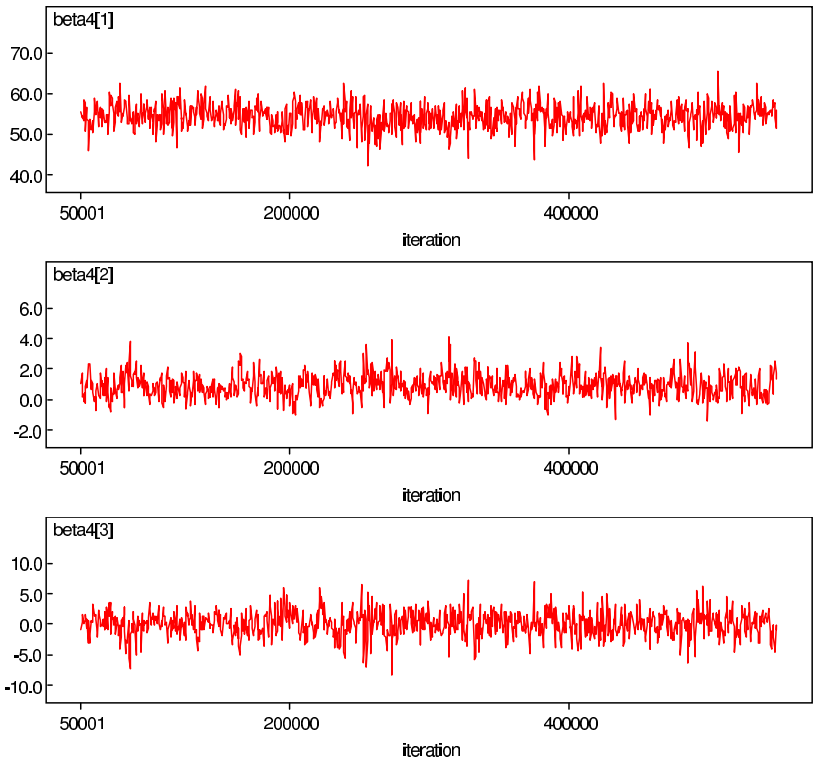
Los parámetros correspondientes a la división II:



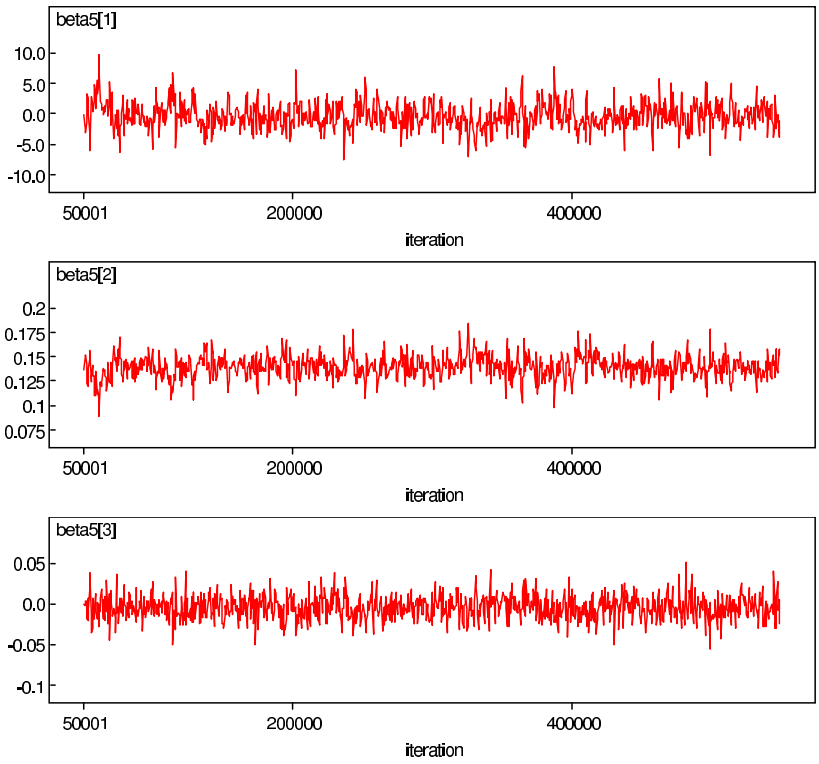
Los parámetros correspondientes a la división III:



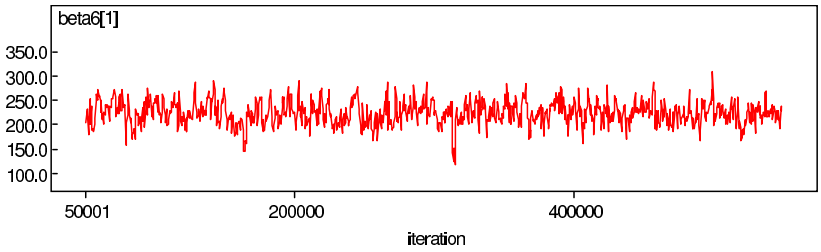
Los parámetros correspondientes a la división IV:

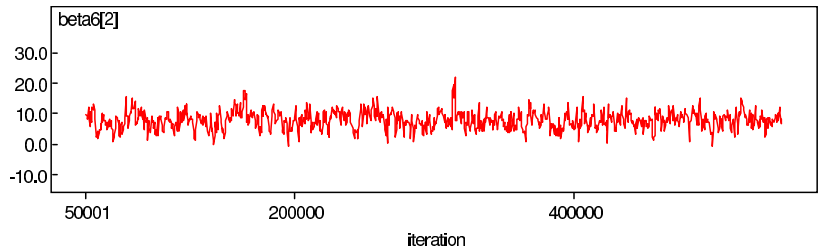


Los parámetros correspondientes a la división V:

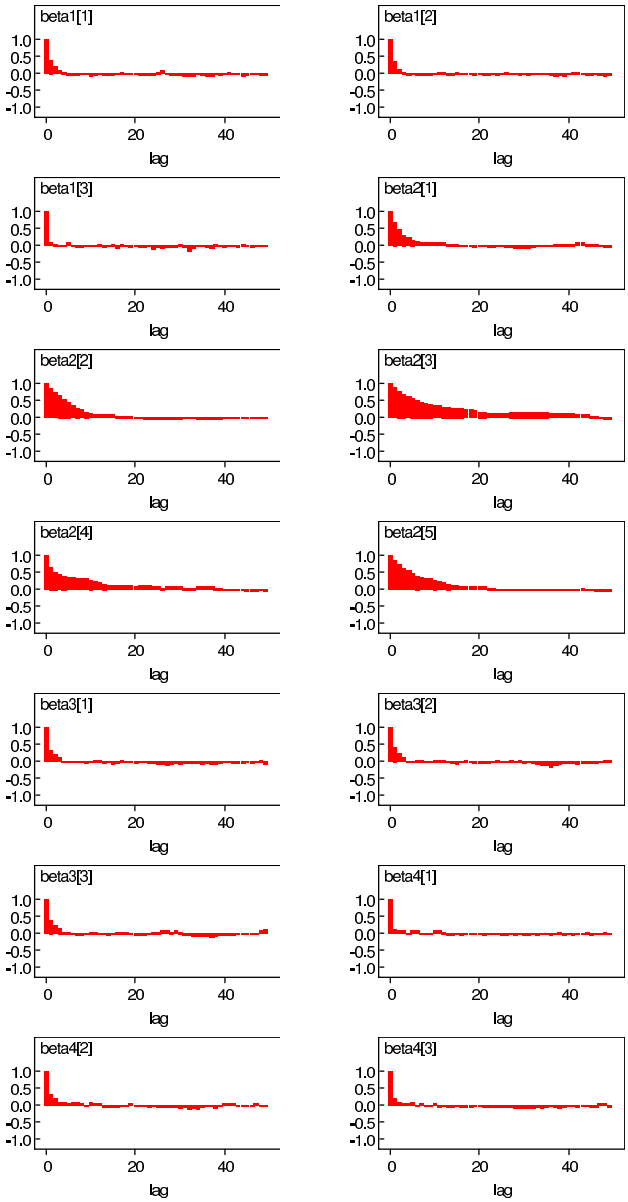


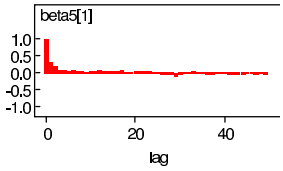
Los parámetros correspondientes a la división VI:

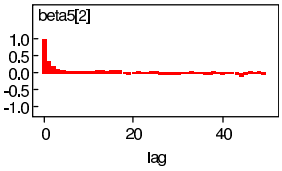


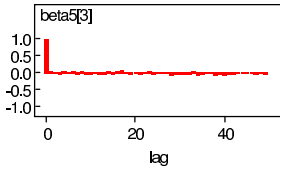


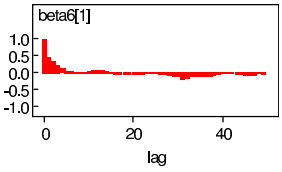
Por su parte la autocorrelación de los parámetros, se muestra en los siguientes gráficos:

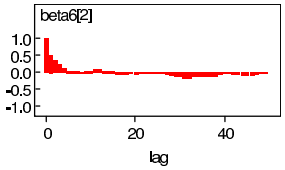



beta5[1]
1.0
0.5
0.0
-0.5
-1.0
0
20
40
lag


beta5[2]
1.0
0.5
0.0
-0.5
-1.0
0
20
40
lag


beta5[3]
1.0
0.5
0.0
-0.5
-1.0
0
20
40
lag


beta6[1]
1.0
0.5
0.0
-0.5
-1.0
0
20
40
lag


beta6[2]
1.0
0.5
0.0
-0.5
-1.0
0
20
40
lag

4.4. Datos faltantes y predicción

Como vimos en la sección anterior, esta aproximación puede ser extendida para datos faltantes del vector de respuesta. Como caso particular, si tenemos datos faltantes para el vector de respuesta para el año 2007, al correr el programa para predecir el valor de la variable respuesta utilizando el modelo SUR, nos da los siguientes resultados:

Parámetro	**Media**	**Desv. est.**	**ECM**	**2.5 %**	**Mediana**	**97.5 %**
y.pred[1]	89.63	2.139	0.08455	85.4	89.55	94.2
y.pred[2]	20.93	1.738	0.08934	17.39	20.97	24.3
y.pred[3]	318.3	12.7	0.5233	291.7	319.1	340
y.pred[4]	59.7	7.101	0.2601	45.17	60.05	73.14
y.pred[5]	31.31	0.9628	0.03847	29.26	31.32	33.02
y.pred[6]	325.7	36.38	1.712	246.4	327.8	390.4

Los datos observados para 2007 fueron:

y.obs[1]	93.737
y.obs[2]	21.883
y.obs[3]	333.406
y.obs[4]	75.077
y.obs[5]	31.521
y.obs[6]	372.180

4.5. Código de WINBUGS

```
model {
for (t in 1:17) {
y[t,1:M] ~ dmnorm(mu[t,1:M],Omega[1:M,1:M])
}

for (t in 1:18) {
# Medias para las observaciones de cada sector
mu[t,1] <- beta1[1] + beta1[2]*I.3[t] + beta1[3]*I.5[t]

mu[t,2] <- beta2[1] + beta2[2]*II.2[t] + beta2[3]*II.3[t] + beta2[4]*II.4[t] + beta2[5]*II.5[t]

mu[t,3] <- beta3[1] + beta3[2]*III.4[t] + beta3[3]*III.5[t]

mu[t,4] <- beta4[1] + beta4[2]*IV.2[t] + beta4[3]*IV.4[t]

mu[t,5] <- beta5[1] + beta5[2]*V.2[t] + beta5[3]*V.6[t]

mu[t,6] <- beta6[1] + beta6[2]*VI.4[t]
}
```

```
# Distribuciones iniciales para los coeficientes de cada regresion
for (j in 1:3) {
beta1[j] ~ dnorm(0,1.05E-6)
}
for (j in 1:5) {
beta2[j] ~ dnorm(0,1.05E-6)
}
for (j in 1:3) {
beta3[j] ~ dnorm(0,1.05E-6)
}
for (j in 1:3) {
beta4[j] ~ dnorm(0,1.05E-6)
}
for (j in 1:3) {
beta5[j] ~ dnorm(0,1.05E-6)
}
for (j in 1:2) {
beta6[j] ~ dnorm(0,1.05E-6)
}

Omega[1 : M , 1 : M]  ~ dwish(R[ , ], 6)

# Prediccion:

y.pred[1:M] ~ dmnorm(mu[18,1:M],Omega[1:M,1:M])

# Matriz de covarianzas cruzadas de dimension M=6
Sigma[1:M,1:M] <- inverse(Omega[1:M, 1:M])

}

Datos 0: Especificacion del modelo
list(T=17,M=6,R = structure(.Data = c(1, 0, 0, 0, 0, 0,
                                      0, 1, 0, 0, 0, 0,
                                      0, 0, 1, 0, 0, 0,
                                      0, 0, 0, 1, 0, 0,
                                      0, 0, 0, 0, 1, 0,
                                      0, 0, 0, 0, 0, 1),.Dim = c(6, 6)))

Datos 0 : Respuestas (todos los sectores)
y[,1]   y[,2]     y[,3]       y[,4]   y[,5]      y[,6]        y[,7]       y[,8]
70.663 15.602453 205.524504 48.040 17.270263 225.0581530  94.872569 158.670333
72.247 15.765214 212.578028 50.385 17.336844 238.7497610  98.124778 166.1254163
70.637 15.963081 221.427423 53.753 17.868653 251.4017233 103.31707  173.74016
72.702 16.257510 219.934043 55.379 18.326503 251.6287198 107.48007  183.208124
```

```
73.373 16.669741 228.891644 60.047 19.200948 268.6960965 116.84213 193.145790
74.005 16.223013 217.581703 45.958 19.613766 226.9599213 111.08117 192.5265
76.646 17.538253 241.151930 50.448 20.511711 237.8590125 120.00070 193.62652
76.791 18.322526 265.113421 55.132 21.580153 263.3132973 131.92273 200.84723
77.397 18.824249 284.642713 57.461 21.979485 278.1614163 140.71588 210.097093
80.196 18.431123 296.631274 60.328 25.456890 286.8183988 151.67593 217.7044
80.641 19.133817 317.091621 62.859 26.216944 321.8385280 165.46885 229.780789
83.456 19.415210 304.990489 59.292 26.817464 318.0354060 171.80594 240.224338
83.506 19.494208 303.003922 60.565 27.077346 318.0793388 174.89942 250.385653
86.194 20.207731 299.156878 62.561 27.481690 322.7322993 183.59106 260.249777
89.152 20.903021 311.013707 66.357 28.250640 340.3793090 200.53687 270.40758
87.324 21.334020 315.314074 68.549 28.743546 349.5180405 214.68667 286.0449783
91.903 21.836000 330.026593 73.501 30.332000 362.3490000 234.18908 301.39580
END


Datos 1: Sector Agropecuario
 I.3[]      I.5[]
14.254687 1.447143
15.508869 1.453276
16.097922 1.246425
16.957980 1.1916
17.482730 1.260019
18.027251 1.404384
17.864740 1.530023
18.658038 1.570586
20.019028 1.233292
21.328433 1.286107
22.222861 1.402938
22.856439 1.520938
23.469864 1.554452
23.755488 1.564966
24.397782 1.515432
24.891685 1.52293
25.679749 1.517898
END

Datos 2: Sector Mineria
II.2[]         II.3[]     II.4[]   II.5[]
0.19667575 0.067566083 1.001978333 101.8833333
0.18604866 0.063111333 1.002238333 94.58333333
0.19398283 0.062532667 1.032061833 95.15
0.20224383 0.071886917 1.114399750 99.99166667
0.19573525 0.070000583 1.165427667 101.9833333
0.20970200  0.073840167 1.275369333 114.3583333
```

```
0.21337900  0.071219500 1.437226333 118.55
0.22728008 0.075882000 1.426829833 123.05
0.24117350 0.075097000 1.379678833 123
0.20662183 0.068652917 1.504071500 117.3333333
0.23105616 0.072504583 1.451388333 121.2333333
0.22789400 0.076796333 1.201756250 122.0416667
0.22081550 0.073248250 1.156312750 113.9833333
0.21140741 0.070522917 1.403141417 114.15
0.20622416 0.070530250 1.452809667 122.3666667
0.21603066 0.077589167 1.548112583 128.7872857
0.20408716 0.073986250 1.597187417 129.7072
END

Datos 3: Sector Manufactura
III.4[] III.5[]
0.446436083  4.779239917
0.467351583  5.0855845
0.505698000  5.386590917
0.542073750  5.846979833
0.583044333  6.495347
0.648263083  7.447647417
0.753708333  8.281070833
0.903527500 10.30605833
1.014006333 12.47587383
1.143240333 14.48369075
1.291231750 16.28788917
1.198941750 17.1292315
1.071209167 16.81820058
1.062104750 17.10723842
1.115229667 17.08182233
1.166249750 17.87254408
1.202134333 18.79552708
END

Datos 4: Sector Construccion
IV.2[]     IV.4[]
1.387714500 1.257094341
2.315167167 1.398185905
2.771702083 1.600812124
3.515010917 1.698601603
4.176881000 1.979469151
2.947853667 1.198991667
3.646622583 1.359925
4.414476833 1.476333333
5.372191750 2.015383333
```

```
6.238960917 2.6726
5.589870500 1.946791667
4.393163667 1.688075
3.819353818 1.683963636
4.863530253 1.972400772
4.928671324 1.991305195
4.995236442 2.010622886
5.163168980 2.059358156
END

Datos 5: Sector Electricidad
V.2[] V.6[]
130.1251115 67.359
130.5887834 74.344
134.2922995 74.77
137.4807669 74.601
143.5704019 57.01
146.4452664 66.158
152.6985606 65.904
159.83     54.605
168.98     62.245
179.07     69.246
190         52.511
194.92     54.363
198.88     41.342
200.94     56.191
205.39     64.299
215.63     75.237
221.9       84.124
END

Datos 6: Sector Comercio
VI.4[]
5.52638
5.95901
6.08479
6.16701
6.363
6.1795
6.75617
7.37619
7.49313
7.2229
8.29503
8.4006
```

```
8.85799
9.36173
10.79555
11.80345
12.17657
END

Valores iniciales  (Minimos cuadrados)
list(beta1=c(0,0,0),
beta2=c(0, 0, 0, 0, 0),
beta3=c(0,0,0),
beta4=c(0,0,0),
beta5=c(0,0,0),
beta6=c(0,0),
Omega = structure(.Data = c(1, 0, 0, 0, 0, 0,
                            0, 1, 0, 0, 0, 0,
                            0, 0, 1, 0, 0, 0,
                            0, 0, 0, 1, 0, 0,
                            0, 0, 0, 0, 1, 0,
                            0, 0, 0, 0, 0, 1),.Dim = c(6, 6)))
```

5. Conclusiones

En este trabajo se analizó un modelo de regresión lineal descrito por primera vez por Zellner (1962), que se escribe como p regresiones lineales múltiples separadas, cuyas variables de respuesta están relacionadas entre sí. Dicho modelo es conocido como SUR, Regresiones Aparentemente no Relacionadas, (cuyas siglas en inglés se leen como *Seemingly Unrelated Regressions*) y lo podemos ver como una extensión de los modelos de regresión lineal que permite errores correlacionados entre las ecuaciones. El modelo SUR también corresponde a un caso particular del modelo de regresión multivariada con una estructura específica en la matriz de varianzas-covarianzas de los vectores de respuesta. Tal vez sea la econometría el área en la que más se ha utilizado este modelo, sin embargo en cualquier regresión multivariada en donde sus variables se relacionen es de gran utilidad.

El modelo SUR fue analizado en este trabajo desde una perspectiva bayesiana y se sugirió un método para aproximar la densidad predictiva del modelo SUR correspondiente a una distribución inicial no informativa a través del muestreo Gibbs. La distribución inicial no informativa utilizada fue la de Jeffreys; sin embargo, se pueden utilizar distribuciones informativas iniciales que sean sencillas para programar.

Dicha aproximación se utilizó posteriormente con datos reales en una aplicación econométrica del modelo SUR, haciendo uso de WINBUGS.

Por los datos analizados podemos ver que, en términos generales las varianzas de las distribuciones finales de los coeficientes del modelo SUR son más pequeñas que las varianzas correspondientes a las regresiones independientes. Esto se debe a que el modelo SUR permite utilizar la información de las otras regresiones al hacer inferencias sobre cada una de las regresiones en particular.

Por otro lado, de acuerdo a lo expuesto en este trabajo se puede decir que el enfoque bayesiano ofrece una gran flexibilidad en la solución de problemas estadísticos ya que permite la incorporación de información previa al momento del análisis, así como la actualización de este conocimiento.

6. Bibliografía

Allison, P. D. (1984) *Event history analysis: Regression for Longitudinal Event Data*. Newbury Park, CA: Sage Publications.

Bernardo, J.M. y Smith, A.F.M. (1994) *Bayesian Theory*. Chichester, Wiley.

Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference. *J. Royal Statistical Society* B 41, 113-147.

Box, G. E. P. y Tiao, G. C. (1973) *Bayesian Inference in Statistical Analysis*. Reading: Addison-Wesley.

Campbell, J.L., Lo. A. W. y Mackinlay, A.C. (1997). *The Econometrics of Financial Markets*. Princeton University Press.

Drèze, J. H. (1977) Bayesian regression analysis using poly-t densities.*New Developments in the Applications of Bayesian Methods* (eds A. Aykac y C. Brumtat), cap. 10. Amsterdam: North-Holland.

Gamerman, D.(1997) *Markov chain Monte Carlo. Stochastic simulation for Bayesian inference* Chapman & Hall

Gelfand, A.E., Hills, S.E., Racine-Poon, A. y Smith, A.F.M. (1989) Illustration of Bayesian inference in normal data models using Gibbs sampling, *Technical Report* Nottingham University.

Gelfand, A.E., y Smith, A.F.M. (1990) Sampling-based approaches to calculating marginal densities, *J. Am. Statis. Ass.* 85, 398-409.

Gelman, A., Carlin J. B. , Stern H. S. y Rubin D. B. (1995) *Bayesian Data Analysis*. New York:Chapman & Hall/CRC.

Geman, S. y Geman, D. (1984) Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images, *IEEE Trans. Pattn Anal. Mach. Intell.* 6, 721-741.

Gilks W.R. , Richardson S. y Spiegelhalter D.J., (1996) *Markov chain Monte Carlo in practice*, ed. Chapman & Hall

Gutiérrez, E.,(1997) Métodos computacionales en la inferencia bayesiana, *Serie Monografías* Vol. 6, No. 15, IIMAS, UNAM

Gutiérrez, E. (1998) Análisis Bayesiano de modelos jerárquicos lineales, *Serie Monografías* Vol. 7, No. 16, IIMAS, UNAM

Little, R.J.A. y Rubin, D.B. (1987), *Statistical Analysis with Missing Data*. New York: Wiley.

Oetiker, T., Partl, H., Hyna, I.y Schlegl, E. (1998) *The Not So Short Introduction to LaTeX2ε or LaTeX2ε in 87 minutes*

Percy, D. (1992) Prediction for Seemingly Unrelated Regressions *J. R. Statist.* **54** No. 1, pp 243-252, University of Liverpool, UK.

Press, S. J. (1972) *Applied Multivariate Analysis.* New York: Holt, Rinehart and Winston.

Spiegelhalter D., Thomas A. y Best N. (1999) *WINBUGS version 1.2 user manual*

Srivastava, V. K. y Giles, D. E. A. (1987) *Seemingly Unrelated Equations Models.* New York: Dekker.

Zellner, A. (1962) An efficient method of estimating seemingly unrelated regression equations and tests for aggregation bias *Journal of the American Statistical Association* 57: 348368.

Zellner, A. (1971). *An introduction to Bayesian Inference in Econometrics.* New York: Wiley.

Zellner, A. y Chetty, V. K. (1965) Prediction and decision problems in regression models from the Bayesian point of view *J. Am. Statist. Ass.*, 60, 608-616.

Proyecto BUGS: *http* : *//www.mrc − bsu.cam.ac.uk/bugs*

Proyecto R: *http* : *//www.r − project.org/*